SemEval-2014 Task 7: Analysis of Clinical Text

Sameer Pradhan[1], **Noémie Elhadad**[2], **Wendy Chapman**[3],
Suresh Manandhar[4] **and Guergana Savova**[1]

[1]Harvard University, Boston, MA, [2]Columbia University, New York, NY
[3]University of Utah, Salt Lake City, UT, [4]University of York, York, UK

{sameer.pradhan,guergana.savova}@childrens.harvard.edu, noemie.elhadad@columbia.edu,
wendy.chapman@utah.edu, suresh@cs.york.ac.uk

Abstract

This paper describes the SemEval-2014, Task 7 on the Analysis of Clinical Text and presents the evaluation results. It focused on two subtasks: (i) identification (Task A) and (ii) normalization (Task B) of diseases and disorders in clinical reports as annotated in the Shared Annotated Resources (ShARe)[1] corpus. This task was a follow-up to the ShARe/CLEF eHealth 2013 shared task, subtasks 1a and 1b,[2] but using a larger test set. A total of 21 teams competed in Task A, and 18 of those also participated in Task B. For Task A, the best system had a strict F_1-score of 81.3, with a precision of 84.3 and recall of 78.6. For Task B, the same group had the best strict accuracy of 74.1. The organizers have made the text corpora, annotations, and evaluation tools available for future research and development at the shared task website.[3]

1 Introduction

A large amount of very useful information—both for medical researchers and patients—is present in the form of unstructured text within the clinical notes and discharge summaries that form a patient's medical history. Adapting and extending natural language processing (NLP) techniques to mine this information can open doors to better, novel, clinical studies on one hand, and help patients understand the contents of their clinical records on the other. Organization of this shared task helps establish state-of-the-art benchmarks and paves the way for further explorations. It tackles two important sub-problems in NLP—named entity recognition and word sense disambiguation. Neither of these problems are new to NLP. Research in general-domain NLP goes back to about two decades. For an overview of the development in the field through roughly 2009, we refer the refer to Nadeau and Sekine (2007). NLP has also penetrated the field of bimedical informatics and has been particularly focused on biomedical literature for over the past decade. Advances in that sub-field has also been documented in surveys such as one by Leaman and Gonzalez (2008). Word sense disambiguation also has a long history in the general NLP domain (Navigli, 2009). In spite of word sense annotations in the biomedical literature, recent work by Savova et al. (2008) highlights the importance of annotating them in clinical notes. This is true for many other clinical and linguistic phenomena as the various characteristics of the clinical narrative present a unique challenge to NLP. Recently various initiatives have led to annotated corpora for clinical NLP research. Probably the first comprehensive annotation performed on a clinical corpora was by Roberts et al. (2009), but unfortunately that corpus is not publicly available owing to privacy regulations. The i2b2 initiative[4] challenges have focused on such topics as concept recognition (Uzuner et al., 2011), coreference resolution (Uzuner et al., 2012), temporal relations (Sun et al., 2013) and their datasets are available to the community. More recently, the Shared Annotated Resources (ShARe)[1] project has created a corpus annotated with disease/disorder mentions in clinical notes as well as normalized them to a concept unique identifier (CUI) within the SNOMED-CT subset of the Unified Medical Language System[5]

[1] http://share.healthnlp.org

[2] https://sites.google.com/site/shareclefehealth/evaluation

[3] http://alt.qcri.org/semeval2014/task7/

[4] http://www.i2b2.org

[5] https://uts.nlm.nih.gov/home.html

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 54–62,
Dublin, Ireland, August 23-24 2014.

	Train	Development	Test
Notes	199	99	133
Words	94K	88K	153K
Disorder mentions	5,816	5,351	7,998
CUI-less mentions	1,639 (28%)	1,750 (32%)	1,930 (24%)
CUI-ied mentions	4,117 (72%)	3,601 (67%)	6,068 (76%)
Contiguous mentions	5,165 (89%)	4,912 (92%)	7,374 (92%)
Discontiguous mentions	651 (11%)	439 (8%)	6,24 (8%)

Table 1: Distribution of data in terms of notes and disorder mentions across the training, development and test sets. The disorders are further split according to two criteria – whether they map to a CUI or whether they are contiguous.

(UMLS) (Campbell et al., 1998). The task of normalization is a combination of word/phrase sense disambiguation and semantic similarity where a phrase is mapped to a unique concept in an ontology (based on the description of that concept in the ontology) after disambiguating potential ambiguous surface words, or phrases. This is especially true with abbreviations and acronyms which are much more common in clinical text (Moon et al., 2012). The SemEval-2014 task 7 was one of nine shared tasks organized at the SemEval-2014. It was designed as a follow up to the shared tasks organized during the ShARe/CLEF eHealth 2013 evaluation (Suominen et al., 2013; Pradhan et al., 2013; Pradhan et al., 2014). Like the previous shared task, we relied on the ShARe corpus, but with more data for training and a new test set. Furthermore, in this task, we provided the options to participants to utilize a large corpus of unlabeled clinical notes. The rest of the paper is organized as follows. Section 2 describes the characteristics of the data used in the task. Section 3 describes the tasks in more detail. Section 4 explains the evaluation criteria for the two tasks. Section 5 lists the participants of the task. Section 6 discusses the results on this task and also compares them with the ShARe/CLEF eHealth 2013 results, and Section 7 concludes.

2 Data

The ShARe corpus comprises annotations over de-identified clinical reports from a US intensive care department (version 2.5 of the MIMIC II database [6]) (Saeed et al., 2002). It consists of discharge summaries, electrocardiogram, echocardiogram, and radiology reports. Access to data was carried out following MIMIC user agreement requirements for access to de-identified medical data. Hence, all participants were required to register for the evaluation, obtain a US human subjects training certificate[7], create an account to the password-protected MIMIC site, specify the purpose of data usage, accept the data use agreement, and get their account approved. The annotation focus was on disorder mentions, their various attributes and normalizations to an UMLS CUI. As such, there were two parts to the annotation: identifying a span of text as a disorder mention and normalizing (or mapping) the span to a UMLS CUI. The UMLS represents over 130 lexicons/thesauri with terms from a variety of languages and integrates resources used world-wide in clinical care, public health, and epidemiology. A disorder mention was defined as any span of text which can be mapped to a concept in SNOMED-CT and which belongs to the Disorder semantic group[8]. It also provided a semantic network in which every concept is represented by its CUI and is semantically typed (Bodenreider and McCray, 2003). A concept was in the Disorder semantic group if it belonged to one of the following 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. The Finding semantic type was left out as it is very noisy and our pilot study showed lower annotation agreement on it. Following are the salient aspects of the guidelines used to

[6] http://mimic.physionet.org – Multiparameter Intelligent Monitoring in Intensive Care

[7] The course was available free of charge on the Internet, for example, via the CITI Collaborative Institutional Training Initiative at https://www.citiprogram.org/Default.asp or, the US National Institutes of Health (NIH) at http://phrp.nihtraining.com/users.

[8] Note that this definition of Disorder semantic group did not include the Findings semantic type, and as such differed from the one of UMLS Semantic Groups, available at http://semanticnetwork.nlm.nih.gov/SemGroups

annotate the data.

- Annotations represent the most specific disorder span. For example, *small bowel obstruction* is preferred over *bowel obstruction*.
- A disorder mention is a concept in the SNOMED-CT portion of the Disorder semantic group.
- Negation and temporal modifiers are not considered part of the disorder mention span.
- All disorder mentions are annotated—even the ones related to a person other than the patient and including acronyms and abbreviations.
- Mentions of disorders that are coreferential/anaphoric are also annotated.

Following are a few examples of disorder mentions from the data.

Patient found to have *lower extremity DVT*. (E1)

In example (E1), lower extremity DVT is marked as the disorder. It corresponds to CUI C0340708 (preferred term: Deep vein thrombosis of lower limb). The span DVT can be mapped to CUI C0149871 (preferred term: Deep Vein Thrombosis), but this mapping would be incorrect because it is part of a more specific disorder in the sentence, namely lower extremity DVT.

A *tumor* was found in the left *ovary*. (E2)

In example (E2), *tumor ... ovary* is annotated as a discontiguous disorder mention. This is the best method of capturing the exact disorder mention in clinical notes and its novelty is in the fact that either such phenomena have not been seen frequently enough in the general domain to gather particular attention, or the lack of a manually curated general domain ontology parallel to the UMLS.

Patient admitted with *low blood pressure*. (E3)

There are some disorders that do not have a representation to a CUI as part of the SNOMED CT within the UMLS. However, if they were deemed important by the annotators then they were annotated as CUI-less mentions. In example (E3), *low blood pressure* is a finding and is normalized as a CUI-less disorder. We constructed the annotation guidelines to require that the disorder be a reasonable synonym of the lexical description of a SNOMED-CT disorder. There are a few instances where the disorders are abbreviated or shortened in the clinical note. One example is w/r/r, which is an abbreviation for concepts wheezing (CUI C0043144), rales (CUI C0034642), and ronchi (CUI C0035508). This abbreviation is also sometimes written as r/w/r and r/r/w. Another is *gsw* for *gunshot wound* and *tachy* for *tachycardia*. More details on the annotation scheme is detailed in the guidelines[9] and in a forthcoming manuscript. The annotations covered about 336K words. Table 1 shows the quantity of the data and the split across the training, development and test sets as well as in terms of the number of notes and the number of words.

2.1 Annotation Quality

Each note in the training and development set was annotated by two professional coders trained for this task, followed by an open adjudication step. By the time we reached annotating the test data, the annotators were quite familiar with the annotation and so, in order to save time, we decided to perform a single annotation pass using a senior annotator. This was followed by a correction pass by the same annotator using a checklist of frequent annotation issues faced earlier. Table 2 shows the inter-annotator agreement (IAA) statistics for the adjudicated data. For the disorders we measure the agreement in terms of the F_1-score as traditional agreement measures such as Cohen's kappa and Krippendorf's alpha are not applicable for measuring agreement for entity mention annotation. We computed agreements between the two annotators as well as between each annotator and the final adjudicated gold standard. The latter is to give a sense of the fraction of corrections made in the process of adjudication. The strict criterion considers two mentions correct if they agree in terms of the class and the exact string, whereas the relaxed criteria considers overlapping strings of the

[9] http://goo.gl/vU8KdW

	Disorder		CUI	
	Relaxed F_1	Strict F_1	Relaxed Acc.	Strict Acc.
A1-A2	90.9	76.9	77.6	84.6
A1-GS	96.8	93.2	95.4	97.3
A2-GS	93.7	82.6	80.6	86.3

Table 2: Inter-annotator (A1 and A2) and gold standard (GS) agreement as F_1-score for the Disorder mentions and their normalization to the UMLS CUI.

Institution	User ID	Team ID
University of Pisa, Italy	attardi	UniPI
University of Lisbon, Portugal	francisco	ULisboa
University of Wisconsin, Milwaukee, USA	ghiasvand	UWM
University of Colorado, Boulder, USA	gung	CLEAR
University of Guadalajara, Mexico	herrera	UG
Taipei Medical University, Taiwan	hjdai	TMU
University of Turku, Finland	kaewphan	UTU
University of Szeged, Hungary	katona	SZTE-NLP
Queensland University of Queensland, Australia	kholghi	QUT_AEHRC
KU Leuven, Belgium	kolomiyets	KUL
Universidade de Aveiro, Portugal	nunes	BioinformaticsUA
University of the Basque Country, Spain	oronoz	IxaMed
IBM, India	parikh	ThinkMiners
easy data intelligence, India	pathak	ezDI
RelAgent Tech Pvt. Ltd., India	ramanan	RelAgent
Universidad Nacional de Colombia, Colombia	riveros	MindLab-UNAL
IIT Patna, India	sikdar	IITP
University of North Texas, USA	solomon	UNT
University of Illinois at Urbana Champaign, USA	upadhya	CogComp
The University of Texas Health Science Center at Houston, USA	wu	UTH_CCB
East China Normal University, China	yi	ECNU

Table 3: Participant organization and the respective User IDs and Team IDs.

same class as correct. The reason for checking the class is as follows. Although we only use the disorder mention in this task, the corpus has been annotated with some other UMLS types as well and therefore there are instances where a different UMLS type is assigned to the same character span in the text by the second annotator. If exact boundaries are not taken into account then the IAA agreement score is in the mid-90s. For the task of normalization to CUIs, we used accuracy to assess agreement. For the relaxed criterion, all overlapping disorder spans with the same CUI were considered correct. For the strict criterion, only disorder spans with identical spans and the same CUI were considered correct.

3 Task Description

The participants were evaluated on the following two tasks:

- **Task A** – Identification of the character spans of disorder mentions.
- **Task B** – Normalizing disorder mentions to SNOMED-CT subset of UMLS CUIs.

For Task A, participants were instructed to develop a system that predicts the spans for disorder mentions. For Tasks B, participants were instructed to develop a system that predicts the UMLS CUI within the SNOMED-CT vocabulary. The input to Task B were the disorder mention predictions from Task A. Task B was optional. System outputs adhered to the annotation format. Each participant was allowed to submit up to three runs. The entire set of unlabeled MIMIC clinical notes (excluding the test notes) were made available to the participants for potential unsupervised approaches to enhance the performance of their systems. They were allowed to use additional annotations in their systems, but this counted towards the total allowable runs; systems that used annotations outside of those provided were evaluated separately. The evaluation for all tasks was conducted using the blind, withheld test data. The participants were provided a training set containing clinical text as well as pre-annotated spans and named entities for disorders (Tasks A and B).

4 Evaluation Criteria

The following evaluation criteria were used:

- **Task A** – The system performance was evaluated against the gold standard using the F_1-score of the Precision and Recall values. There were two variations: (i) Strict; and (ii) Relaxed. The formulae for computing these metrics are mentioned below.

$$Precision = P = \frac{D_{tp}}{D_{tp} + D_{fp}} \quad (1)$$

$$Recall = R = \frac{D_{tp}}{D_{tp} + D_{fn}} \quad (2)$$

Where, D_{tp} = Number of true positives disorder mentions; D_{fp} = Number of false positives disorder mentions; D_{fn} = Number of false negative disorder mentions. In the strict case, a span was counted as correct if it was identical to the gold standard span, whereas

Team ID	User ID	Run	Task A Strict P (%)	Strict R (%)	Strict F_1 (%)	Relaxed P (%)	Relaxed R (%)	Relaxed F_1 (%)	Data
UTH_CCB	wu	0	84.3	78.6	81.3	93.6	86.6	90.0	T+D
UTH_CCB	wu	1	80.8	80.5	80.6	91.6	90.7	91.1	T+D
UTU	kaewphan	1	76.5	76.7	76.6	88.6	89.9	89.3	T+D
UWM	ghiasvand	0	78.7	72.6	75.5	91.1	85.6	88.3	T+D
UTH_CCB	wu	2	68.0	84.9	75.5	83.8	93.5	88.4	T+D
UTU	kaewphan	0	77.3	72.4	74.8	90.1	85.6	87.8	T
IxaMed	oronoz	1	68.1	78.6	73.0	87.2	89.0	88.1	T+D
UWM	ghiasvand	0	77.5	67.9	72.4	90.9	81.2	85.8	T
RelAgent	ramanan	0	74.1	70.1	72.0	89.5	84.0	86.7	T+D
IxaMed	oronoz	0	72.9	70.1	71.5	88.5	80.8	84.5	T+D
ezDI	pathak	1	75.0	68.2	71.4	91.5	82.7	86.9	T
CLEAR	gung	0	80.7	63.6	71.2	92.0	72.3	81.0	T
ezDI	pathak	0	75.0	67.7	71.2	91.4	81.9	86.4	T
ULisboa	francisco	0	75.3	66.3	70.5	91.4	81.5	86.2	T
ULisboa	francisco	1	75.2	66.0	70.3	90.9	80.6	85.5	T
ULisboa	francisco	2	75.2	66.0	70.3	90.9	80.6	85.5	T
BioinformaticsUA	nunes	0	81.3	60.5	69.4	92.9	69.3	79.4	T+D
ThinkMiners	parikh	0	73.4	65.0	68.9	89.2	80.2	84.4	T
ThinkMiners	parikh	1	74.9	61.7	67.7	90.7	75.8	82.6	T
ECNU	yi	0	75.4	61.1	67.5	89.8	72.2	80.0	T+D
UniPI	attardi	2	71.2	60.1	65.2	89.7	76.6	82.6	T+D
UNT	solomon	0	64.7	62.8	63.8	81.5	79.9	80.7	T+D
UniPI	attardi	1	65.9	61.2	63.5	90.2	77.5	83.4	T+D
BioinformaticsUA	nunes	2	75.3	53.8	62.8	86.5	62.1	72.3	T+D
BioinformaticsUA	nunes	1	60.0	62.1	61.0	69.8	72.3	71.0	T+D
UniPI	attardi	0	53.9	68.4	60.2	77.8	88.5	82.8	T+D
CogComp	upadhya	1	63.9	52.9	57.9	82.3	68.3	74.6	T+D
CogComp	upadhya	2	64.1	52.0	57.4	82.9	67.5	74.4	T+D
CogComp	upadhya	0	63.6	51.5	56.9	81.9	66.5	73.4	T+D
TMU	hjdai	0	52.4	57.6	54.9	91.4	76.5	83.3	T+D
MindLab-UNAL	riveros	2	56.1	53.4	54.7	76.9	67.7	72.0	T
MindLab-UNAL	riveros	1	57.8	51.5	54.5	77.7	65.4	71.0	T
TMU	hjdai	1	62.2	42.9	50.8	89.9	65.2	75.6	T+D
IITP	sikdar	0	50.0	47.9	48.9	81.5	79.7	80.6	T+D
IITP	sikdar	1	47.3	45.8	46.5	78.9	77.6	78.2	T+D
IITP	sikdar	2	45.0	48.1	46.5	76.9	82.6	79.6	T+D
MindLab-UNAL	riveros	0	32.1	56.5	40.9	43.9	72.5	54.7	T
SZTE-NLP	katona	1	54.7	25.2	34.5	88.4	40.1	55.1	T
SZTE-NLP	katona	2	54.7	25.2	34.5	88.4	40.1	55.1	T
QUT_AEHRC	kholghi	0	38.7	29.8	33.7	90.6	70.9	79.5	T+D
SZTE-NLP	katona	0	57.1	20.5	30.2	91.8	32.5	48.0	T
KUL	kolomiyets	0	65.5	17.8	28.0	72.1	19.6	30.8	P
UG	herrera	0	11.4	23.4	15.3	25.9	49.0	33.9	P

Table 4: Performance on *test* data for participating systems on Task A – Identification of disorder mentions.

Team ID	User ID	Run	Task A Strict P (%)	Strict R (%)	Strict F_1 (%)	Relaxed P (%)	Relaxed R (%)	Relaxed F_1 (%)	Data
hjdai	TMU	1	0.687	0.922	0.787	0.952	1.000	0.975	T
wu	UTH_CCB	0	0.877	0.710	0.785	0.962	0.789	0.867	T
wu	UTH_CCB	1	0.828	0.747	0.785	0.941	0.853	0.895	T
Best ShARe/CLEF-2013 performance			**0.800**	**0.706**	**0.750**	**0.925**	**0.827**	**0.873**	T
ghiasvand	UWM	0	0.827	0.675	0.743	0.958	0.799	0.871	T
pathak	ezDI	0	0.813	0.670	0.734	0.954	0.800	0.870	T
pathak	ezDI	1	0.809	0.667	0.732	0.954	0.801	0.871	T
wu	UTH_CCB	2	0.657	0.790	0.717	0.806	0.893	0.847	T
francisco	ULisboa	1	0.803	0.646	0.716	0.954	0.781	0.858	T
francisco	ULisboa	2	0.803	0.646	0.716	0.954	0.781	0.858	T
francisco	ULisboa	0	0.796	0.642	0.711	0.959	0.793	0.868	T
oronoz	IxaMed	0	0.766	0.650	0.703	0.936	0.752	0.834	T
oronoz	IxaMed	1	0.660	0.721	0.689	0.899	0.842	0.870	T
hjdai	TMU	0	0.667	0.414	0.511	0.912	0.591	0.717	T
sikdar	IITP	0	0.525	0.430	0.473	0.862	0.726	0.788	T
sikdar	IITP	2	0.467	0.440	0.453	0.812	0.775	0.793	T
sikdar	IITP	1	0.493	0.410	0.448	0.828	0.706	0.762	T

Table 5: Performance on *development* data for participating systems on Task A – Identification of disorder mentions.

in the relaxed case, a span overlapping with the gold standard span was also considered correct.

- **Task B** – Accuracy was used as the performance measure for Task 1b. It was defined as follows:

$$Accuracy_{strict} = \frac{D_{tp} \cap N_{correct}}{T_g} \quad (3)$$

$$Accuracy_{relaxed} = \frac{D_{tp} \cap N_{correct}}{D_{tp}} \quad (4)$$

Where, D_{tp} = Number of true positive disorder mentions with identical spans as in the gold standard; $N_{correct}$ = Number of correctly normalized disorder mentions; and T_g = Total number of disorder mentions in the gold standard. For Task B, the systems were only evaluated on annotations they identified in Task A. Relaxed accuracy only measured the ability to normalize correct spans. Therefore, it was possible to obtain very high values for this measure by simply dropping any mention with a low confidence span.

5 Participants

A total of 21 participants from across the world participated in Task A and out of them 18 also participated in Task B. Unfortunately, although interested, the ThinkMiners team (Parikh et al., 2014) could not participate in Task B owing to some UMLS licensing issues. The participating organizations along with the contact user's User ID and their chosen Team ID are mentioned in Table 3. Eight teams submitted three runs, six submitted two runs and seven submitted just one run. Out of these, only 13 submitted system description papers. We based our analysis on those system descriptions.

6 System Results

Tables 4 and 6 show the performance of the systems on Tasks A and B. None of the systems used any additional annotated data so we did not have to compare them separately. Both tables mention performance of all the different runs that the systems submitted. Given the many variables, we deliberately left the decision on how many and how to define these runs to the individual participant. They used various different ways to differentiate their runs. Some, for example, UTU (Kaewphan et al., 2014), did it based on the composition of training data, i.e., whether they used just the training data or both the training and the development data for training the final system, which highlighted the fact that adding development data to training bumped the F_1-score on Task A by about 2 percent points. Some participants, however, did not make use of the development data in training their systems. This was partially due to the fact that we had not explicitly mentioned in the task description that participants were allowed to use the development data for training their final models. In order to be fair, we allowed some users an opportunity to submit runs post evaluation where they used the exact same system that they used for evaluation but used the development data as well. We added a column to the results tables showing whether the participant used only the training data (T) or both training and development data (T+D) for training their system. It can be seen that even though the addition of development data helps, there are still systems that perform in the lower percentile who have used both training and development data for training, indicating that both the features and the machine learning classifier contribute to the models. A novel aspect of the SemEval-2014 shared task that differentiates it from the ShARE/CLEF task—other than the fact that it used more data and a new test set—is the fact that SemEval-2014 allowed the use of a much larger set of unlabeled MIMIC notes to inform the models. Surprisingly, only two of the systems (ULisboa (Leal et al., 2014) and UniPi (Attardi et al., 2014)) used the unlabeled MIMIC corpus to generalize the lexical features. Another team—UTH_CCB(Zhang et al., 2014)—used off-the-shelf Brown clusters[10] as opposed to training them on the unlabeled MIMIC II data. For Task B, the accuracy of a system using the *strict* metric was positively correlated with its recall on the disorder mentions that were input to it (i.e., recall for Task A), and did not get penalized for lower precision. Therefore one could essentially gain higher accuracy in Task B by tuning a system to provide the highest mention recall in Task A potentially at the cost of precision and the overall F_1-score and using those mentions as input for Task B. This can be seen from the fact that the run 2 for UTH_CCB (Zhang et al., 2014) system with the lowest F_1-score has

[10]Personal conversation with the participants as it was not very clear in the system description paper.

Team ID	User ID	Task B Run	Strict Acc. (%)	Relaxed Acc. (%)	Data
UTH_CCB	wu	2	74.1	87.3	T+D
UTH_CCB	wu	1	70.8	88.0	T+D
UTH_CCB	wu	0	69.4	88.3	T+D
UWM	ghiasvand	0	66.0	90.9	T+D
RelAgent	ramanan	0	63.9	91.2	T+D
UWM	ghiasvand	0	61.7	90.8	T
IxaMed	oronoz	0	60.4	86.2	T+D
UTU	kaewphan	1	60.1	78.3	T+D
ezDI	pathak	1	59.9	87.8	T
ezDI	pathak	0	59.2	87.4	T
UTU	kaewphan	0	57.7	79.7	T
BioinformaticsUA	nunes	1	53.1	85.5	T+D
BioinformaticsUA	nunes	0	52.7	87.0	T+D
CLEAR	gung	0	52.5	82.5	T
TMU	hjdai	0	48.9	84.9	T+D
UNT	solomon	0	47.0	74.8	T+D
UniPI	attardi	0	46.7	68.3	T+D
BioinformaticsUA	nunes	2	46.3	86.1	T+D
MindLab-UNAL	riveros	2	46.1	86.3	T
IxaMed	oronoz	1	43.9	55.8	T+D
MindLab-UNAL	riveros	0	43.5	77.1	T
UniPI	attardi	1	42.8	69.9	T+D
UniPI	attardi	2	41.7	69.3	T+D
MindLab-UNAL	riveros	1	41.1	79.7	T
ULisboa	francisco	2	40.5	61.5	T
ULisboa	francisco	1	40.4	61.2	T
ULisboa	francisco	0	40.2	60.6	T
ECNU	yi	0	36.4	59.5	T+D
TMU	hjdai	1	35.8	83.4	T+D
IITP	sikdar	0	33.3	69.6	T+D
IITP	sikdar	2	33.2	69.1	T+D
IITP	sikdar	1	31.9	69.6	T+D
CogComp	upadhya	1	25.3	47.9	T+D
CogComp	upadhya	2	24.8	47.7	T+D
CogComp	upadhya	0	24.4	47.3	T+D
KUL	kolomiyets	0	16.5	92.8	P
UG	herrera	0	12.5	53.4	P

Table 6: Performance on *test* data for participating systems on Task B – Normalization of disorder mentions to UMLS (SNOMED-CT subset) CUIs.

Team ID	User ID	Task B Run	Strict Acc. (%)	Relaxed Acc. (%)	Data
TMU	hjdai	0	0.716	0.777	T
TMU	hjdai	1	0.716	0.777	T
UTH_CCB	wu	2	0.713	0.903	T
UTH_CCB	wu	1	0.680	0.910	T
UTH_CCB	wu	0	0.647	0.910	T
UWM	ghiasvand	0	0.623	0.923	T
ezDI	pathak	0	0.603	0.900	T
ezDI	pathak	1	0.600	0.899	T
Best ShARe/CLEF-2013 performance			**0.589**	**0.895**	T
IxaMed	oronoz	0	0.556	0.855	T
IxaMed	oronoz	1	0.421	0.584	T
ULisboa	francisco	2	0.388	0.601	T
ULisboa	francisco	1	0.385	0.596	T
ULisboa	francisco	0	0.377	0.588	T
IITP	sikdar	2	0.318	0.724	T
IITP	sikdar	0	0.312	0.725	T
IITP	sikdar	1	0.299	0.730	T

Table 7: Performance on *development* data for some participating systems on Task B – Normalization of disorder mentions to UMLS (SNOMED-CT subset) CUIs.

the best accuracy for Task B and vice-versa for run 0 with run 1 in between the two. In order to fairly compare the performance between two systems one would have to provide perfect mentions as input to Task B. One of the systems—UWM Ghiasvand and Kate (2014)—did run some ablation experiments using gold standard mentions as input to Task B and obtained a best performance of 89.5F_1-score (Table 5 of Ghiasvand and Kate (2014)) as opposed to 62.3 F_1-score (Table 7) in the more realistic setting which is a huge difference. In the upcoming SemEval-2014 where this same evaluation is going to carried out under Task 14, we plan to perform supplementary evaluation where gold disorder mentions would be input to the system while attempting Task B. An interesting outcome of planning a follow-on evaluation to the ShARe/CLEF eHealth 2013 task was that we could, and did, use the test data from the ShARe/CLEF eHealth 2013 task as the development set for this evaluation. After the main evaluation we asked participants to provide the system performance on the development set using the same number and run convention that they submitted for the main evaluation. These results are presented in Tables 5 and 7. We have inserted the best performing system score from the ShARe/CLEF eHealth 2013 task in these tables. For Task A, referring to Tables 4 and 5, there is a boost of 3.7 absolute percent points for the F_1-score over the same task (Task 1a) in the ShARe/CLEF eHealth 2013. For Task B, referring to Tables 6 and 7, there is a boost of 13.7 percent points for the F_1-score over the same task (Task 1b) in the ShARe/CLEF eHealth 2013 evaluation. The participants used various approaches for tackling the tasks, ranging from purely rule-based/unsupervised (RelAgent (Ramanan and Nathan, 2014), (Matos et al., 2014), KUL[11]) to a hybrid of rules and machine learning classifiers. The top performing systems typically used the latter. Various versions of the IOB formulation were used for tagging the disorder mentions. None of the standard variations on the IOB formulation were explicitly designed or used to handle discontiguous mentions. Some systems used novel variations on this approach. Probably the simplest variation was applied by the UWM team (Ghiasvand and Kate, 2014). In this formulation the following labeled sequence "the/O left/B atrium/I is/O moderately/O

[11] Personal communication with participant.

dilated/I" can be used to represent the discontiguous mention *left atrium...dilated*, and can be constructed as such from the output of the classification. The most complex variation was the one used by the UTH_CCB team (Zhang et al., 2014) where they used the following set of tags—B, I, O, DB, DI, HB, HI. This variation encodes discontiguous mentions by adding four more tags to the I, O and B tags. These are variations of the B and I tags with either a D or a H prefix. The prefix H indicates that the word or word sequence is the shared head, and the prefix D indicates otherwise. Another intermediate approach used by the ULisboa team (Leal et al., 2014) with the tagset—S, B, I, O, E and N. Here, S represents the single token entity to be recognized, E represents the end of an entity (which is part of one of the prior IOB variations) and an N tag to identify non-contiguous mentions. They don't provide an explicit example usage of this tag set in their paper. Yet another variation was used by the SZTE-NLP team (Katona and Farkas, 2014). This used tags B, I, L, O and U. Here, L is used for the last token similar to E earlier, and U is used for a unit-token mention, similar to S earlier. We believe that the only approach that can distinguish between discontiguous disorders that share the same head word/phrase is the one used by the UTH_CCB team (Zhang et al., 2014). The participants used various machine learning classifiers such as MaxEnt, SVM, CRF in combination with rich syntactic and semantic features to capture the disorder mentions. As mentioned earlier, a few participants used the available unlabeled data and also off-the-shelf clusters to better generalize features. The use of vector space models such as cosine similarities as well as continuous distributed word vector representations was useful in the normalization task. They also availed of tools such as MetaMap and cTakes to generate features as well as candidate CUIs during normalizations.

7 Conclusion

We have created a reference standard with high inter-annotator agreement and evaluated systems on the task of identification and normalization of diseases and disorders appearing in clinical reports. The results have demonstrated that an NLP system can complete this task with reasonably high accuracy. We plan to annotate another evaluation using the same data as part of the in the SemEval-2015, Task 14[12] adding another task of template filling where the systems will identify and normalize ten attributes the identified disease/disorder mentions.

Acknowledgments

We greatly appreciate the hard work and feedback of our program committee members and annotators David Harris, Jennifer Green and Glenn Zaramba. Danielle Mowery, Sumithra Velupillai and Brett South for helping prepare the manuscript by summarizing the approaches used by various systems. This shared task was partially supported by Shared Annotated Resources (ShARe) project NIH 5R01GM090187 and Temporal Histories of Your Medical Events (THYME) project (NIH R01LM010090 and U54LM008748).

References

Giuseppe Attardi, Vitoria Cozza, and Daniele Sartiano. 2014. UniPi: Recognition of mentions of disorders in clinical text. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

Olivier Bodenreider and Alexa McCray. 2003. Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36:414–432.

Keith E. Campbell, Diane E. Oliver, and Edward H. Shortliffe. 1998. The Unified Medical Language System: Towards a collaborative approach for solving terminologic problems. *J Am Med Inform Assoc*, 5(1):12–16.

Omid Ghiasvand and Rohit J. Kate. 2014. UWM: Disorder mention extraction from clinical text using crfs and normalization using learned edit distance patterns. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

Suwisa Kaewphan, Kai Hakaka1, and Filip Ginter. 2014. UTU: Disease mention recognition and normalization with crfs and vector space representations. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

Melinda Katona and Richárd Farkas. 2014. SZTE-NLP: Clinical text analysis with named entity recognition. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

André Leal, Diogo Gonçalves, Bruno Martins, and Francisco M. Couto. 2014. ULisboa: Identification and classification of medical concepts. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

[12] http://alt.qcri.org/semeval2015/task14

Robert Leaman and Graciela Gonzalez. 2008. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663.

Sérgio Matos, Tiago Nunes, and José Luís Oliveira. 2014. BioinformaticsUA: Concept recognition in clinical narratives using a modular and highly efficient text processing framework. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

Sungrim Moon, Serguei Pakhomov, and Genevieve B Melton. 2012. Automated disambiguation of acronyms and abbreviations in clinical texts: Window and training size considerations. In *AMIA Annu Symp Proc*, pages 1310–1319.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Roberto Navigli. 2009. Word sense disambiguation. *ACM Computing Surveys*, 41(2):1–69, February.

Ankur Parikh, Avinesh PVS, Joy Mustafi, Lalit Agarwalla, and Ashish Mungi. 2014. ThinkMiners: SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

Sameer Pradhan, Noémie Elhadad, Brett South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *Working Notes of CLEF eHealth Evaluation Labs*.

Sameer Pradhan, Noémie Elhadad, Brett South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova. 2014. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. In *Journal of the American Medical Informatics Association (to appear)*.

S. V. Ramanan and P. Senthil Nathan. 2014. RelAgent: Entity detection and normalization for diseases in clinical records: a linguistically driven approach. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *J Biomed Inform*, 42(5):950–66.

Mohammed Saeed, C. Lieu, G. Raber, and R.G. Mark. 2002. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, 29.

Guergana K. Savova, A. R. Coden, I. L. Sominsky, R. Johnson, P. V. Ogren, P. C. de Groen, and C. G. Chute. 2008. Word sense disambiguation across two domains: Biomedical literature and clinical notes. *J Biomed Inform*, 41(6):1088–1100, December.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–13.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. 2013. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *Working Notes of CLEF eHealth Evaluation Labs*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Özlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of American Medical Informatics Association*, 19(5):786–791, September.

Yaoyun Zhang, Jingqi Wang, Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, and Hua Xu. 2014. UTH_CCB: A report for SemEval 2014 task 7 analysis of clinical text. In *Proceedings of the International Workshop on Semantic Evaluations*, Dublin, Ireland, August.