

# Task 1: ShARe/CLEF eHealth Evaluation Lab 2013

Sameer Pradhan<sup>1</sup>, Noemie Elhadad<sup>2</sup>, Brett R. South<sup>3</sup>, David Martinez<sup>5</sup>, Lee Christensen<sup>3</sup>, Amy Vogel<sup>2</sup>, Hanna Suominen<sup>4</sup>, Wendy W. Chapman<sup>6</sup>, and Guergana Savova<sup>1</sup> \*

<sup>1</sup> Boston Children's Hospital and Harvard Medical School,  
{sameer.pradhan, guergana.savova}@childrens.harvard.edu

<sup>2</sup> Columbia University, noemie@dbmi.columbia.edu, amy.vogel@gmail.com

<sup>3</sup> University of Utah, brett.south@hsc.utah.edu, leenlp@q.com

<sup>4</sup> NICTA and The Australian National University, Australia  
hanna.suominen@nicta.com.au

<sup>5</sup> NICTA and The University of Melbourne, Australia  
david.martinez@nicta.com.au

<sup>6</sup> University of California San Diego wwchapman@ucsd.edu

**Abstract.** This report outlines the Task 1 of the ShARe/CLEF eHealth evaluation lab pilot. This task focused on identification (1a) and normalization (1b) of diseases and disorders in clinical reports. It used annotations from the ShARe corpus. A total of 22 teams competed in Task 1a and 17 of them also participated Task 1b. The best systems had an F1 score of 0.75 (0.80 Precision, 0.71 Recall) in Task 1a and an accuracy of 0.59 in Task 1b. The organizers have made the text corpora, annotations, and evaluation tools available for future research and development.

**Keywords:** Natural Language Processing, Text Normalization, Medical Informatics, Reference Standard Generation

## 1 Introduction

A large amount of very useful information – both for the medical researchers and the 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 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 baselines and paves way to further explorations. The shared task was one of three shared tasks organized at the CLEF eHealth Evaluation Labs [1, 2]

---

\* WWC, BRS, and DLM led the task; WWC, BRS, DLM, NE, SP, and GS defined the task; GS and NE led the annotation effort; AV provided coordination and management of the annotations; HS co-chaired the lab; DLM, BRS and LC processed and distributed the dataset; and DM and WWC led result evaluations

## 2 Data

The ShARe corpus<sup>7</sup> comprises of annotations over de-identified clinical reports from from US intensive care (version 2.5 of the MIMIC II database<sup>8</sup>.) The corpus consisted of discharge summaries and electrocardiogram, echocardiogram, and radiology reports. Although the clinical reports were de-identified, they still needed to be treated with appropriate care and respect. Hence, all participants were required to register to the lab, obtain a US human subjects training certificate<sup>9</sup>, create an account to a password-protected site on the Internet, specify the purpose of data usage, accept the data use agreement, and get their account approved.

## 3 Task Description

The annotation of disorder mentions in clinical reports was carried out as part of the ongoing ShARe project<sup>10</sup>. For this task in the evaluation lab, the focus was on the annotation of disorder mentions only. As such, there were two parts to the annotation: identifying a span of text as a disorder mention and mapping the span to a UMLS [3] CUI. Each note was annotated by two professional coders trained for this task, followed by an open adjudication step. UMLS<sup>11</sup> represented over 130 lexicons/thesauri with terms from a variety of languages. It integrated resources used world-wide in clinical care, public health, and epidemiology. It also provided a semantic network in which every concept is represented by its CUI and is semantically typed [4]. 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<sup>12</sup>. 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 annotations covered about 181,000 words.

## 4 Evaluation Methods

The following evaluation criteria were used:

<sup>7</sup> <https://www.clinicalnlpannotation.org>

<sup>8</sup> Multiparameter Intelligent Monitoring in Intensive Care <http://mimic.physionet.org>

<sup>9</sup> 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/login.php>.

<sup>10</sup> <https://www.clinicalnlpannotation.org>

<sup>11</sup> <https://uts.nlm.nih.gov/home.html>

<sup>12</sup> 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>

- 1a. correctness in identification of the character spans of disorders,
- 1b. correctness in mapping disorders to SNOMED-CT codes,

In Tasks 1a and 1b each participating team was permitted to upload the outputs of up to two systems. Task 1b was optional for Task 1 participants. Teams were allowed to use additional annotations in their systems, but this counted towards the permitted systems; 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 1a and 1b). For Task 1a, participants were instructed to develop a system that predicts the spans for disorder named entities. For Tasks 1b, participants were instructed to develop a system that predicts the SNOMED-CT code. The outputs needed to follow the annotation format. The corpus of reports was split into 200 training and 100 testing. The system performance was evaluated against the criteria by using the F1 score in Task 1a and Accuracy in Tasks 1b. We relied on non-parametric statistical significance tests called random shuffling [5] to better compare the measure values for the systems and benchmarks. In Task 1a, the F1 score was defined as the harmonic mean of Precision (P) and Recall (R); P as  $n_{TP}/(n_{TP} + n_{FP})$ ; R as  $n_{TP}/(n_{TP} + n_{FN})$ ;  $n_{TP}$  as the number of instances, where the spans identified by the system and gold standard were the same;  $n_{FP}$  as the number of spurious spans by the system; and  $n_{FN}$  as the number of missing spans by the system. We referred to the Exact (Relaxed) F1-score if the system span is identical to (overlaps) the gold standard span. In Tasks 1b the Accuracy was defined as the number of pre-annotated spans with correctly generated code divided by the total number of pre-annotated spans. In both tasks, the Exact Accuracy and Relaxed Accuracy were measured. In the Exact Accuracy for Task 1b, *total* was defined as the total number of gold standard named entities. In this case, the system was penalised for incorrect code assignment for annotations that were not detected by the system. In the Relaxed Accuracy for Task 1b, *total* was defined as the total number of named entities with strictly correct span generated by the system. In this case, the system was only evaluated on annotations that were detected by the system.

## 5 System Results

A total of 22 teams competed in Task 1a and 16 of them also participated Task 1b. The performance of these systems is detailed in Tables 1 and 2. The best systems had an F1 score of 0.75 (0.80 Precision, 0.71 Recall) in Task 1a and an accuracy of 0.59 in Task 1b.

## 6 Discussion

We have created a reference standard with high inter-annotator agreement and evaluated systems on the task of identification and normalization of diseases and

**Table 1.** Evaluation for Task 1a. For the column of Strict F1 score, “\*” indicates that the F1 score of the system was significantly better than the one immediately below (random shuffling,  $p < 0.01$ ).

System ID ({team}.{system})	Strict Evaluation			Relaxed Evaluation		
	Precision	Recall	F1 score	Precision	Recall	F1 score
<i>No additional annotations:</i>						
(UTHealthCCB.A).2	0.800	0.706	0.750*	0.925	0.827	0.873
(UTHealthCCB.A).1	0.831	0.663	0.737*	0.954	0.774	0.854
NCBI.1	0.768	0.654	0.707*	0.910	0.796	0.849
NCBI.2	0.757	0.658	0.704*	0.904	0.805	0.852
CLEAR.2	0.764	0.624	0.687*	0.929	0.759	0.836
(Mayo.A).1	0.800	0.573	0.668*	0.936	0.680	0.787
(UCDCSI.A).1	0.745	0.587	0.656	0.922	0.758	0.832
CLEAR.1	0.755	0.573	0.651*	0.937	0.705	0.804
(Mayo.B).1	0.697	0.574	0.629*	0.939	0.766	0.844
CORAL.2	0.796	0.487	0.604	0.909	0.554	0.688
HealthLanguageLABS.1	0.686	0.539	0.604*	0.912	0.701	0.793
LIMSI.2	0.814	0.473	0.598*	0.964	0.563	0.711
LIMSI.1	0.805	0.466	0.590	0.962	0.560	0.708
(AEHRC.A).2	0.613	0.566	0.589*	0.886	0.785	0.833
(WVU.DG&VJ).1	0.614	0.505	0.554*	0.885	0.731	0.801
(WVU.SS&VJ).1	0.575	0.496	0.533	0.848	0.741	0.791
CORAL.1	0.584	0.446	0.505	0.942	0.601	0.734
NIL-UCM.2	0.617	0.426	0.504	0.809	0.558	0.660
KPSCML.2	0.494	0.512	0.503*	0.680	0.687	0.684
NIL-UCM.1	0.621	0.416	0.498	0.812	0.543	0.651
KPSCML.1	0.462	0.523	0.491*	0.651	0.712	0.680
(AEHRC.A).1	0.699	0.212	0.325*	0.903	0.275	0.422
(WVU.AJ&VJ).1	0.230	0.318	0.267*	0.788	0.814	0.801
UCDCSI.2	0.268	0.175	0.212*	0.512	0.339	0.408
SNUBME.2	0.191	0.137	0.160*	0.381	0.271	0.317
SNUBME.1	0.302	0.026	0.047	0.504	0.043	0.079
(WVU.FP&VJ).1	0.024	0.446	0.046	0.088	0.997	0.161
<i>Additional annotations:</i>						
(UCSC.CW&RA).2	0.732	0.621	0.672	0.883	0.742	0.806
(UCSC.CW&RA).1	0.730	0.615	0.668*	0.887	0.739	0.806
RelAgent.2	0.651	0.494	0.562*	0.901	0.686	0.779
RelAgent.1	0.649	0.450	0.532	0.913	0.636	0.750
(WVU.AL&VJ).1	0.492	0.558	0.523*	0.740	0.840	0.787
(THCIB.A).1	0.445	0.551	0.492*	0.720	0.713	0.716
(WVU.RK&VJ).1	0.397	0.465	0.428	0.717	0.814	0.762

**Table 2.** Evaluation for Tasks 1b. For the column of Strict Accuracy, “\*” indicates that the Accuracy of the system was significantly better than the one immediately below (random shuffling,  $p < 0.01$ ). The CORAL systems for Task 1b were not in the results announced on May 14 due to a missing registration until June 17.

System ID ({team}·{system})	Strict Accuracy	Relaxed Accuracy
<i>Task 1b, no additional annotations:</i>		
NCBI.2	0.589*	0.895
NCBI.1	0.587*	0.897
(Mayo.A).2	0.546*	0.860
(UTHealthCCB.A).1	0.514*	0.728
(UTHealthCCB.A).2	0.506	0.717
(Mayo.A).1	0.502*	0.870
KPSCMI.1	0.443*	0.865
CLEAR.2	0.440*	0.704
CORAL.2	0.439*	0.902
CORAL.1	0.410*	0.921
CLEAR.1	0.409*	0.713
NIL-UCM.2	0.362	0.850
NIL-UCM.1	0.362*	0.871
(AEHRC.A).2	0.313*	0.552
(WVU.SS&VJ).1	0.309	0.622
(UCDCSI.B).1	0.299*	0.509
(WVU.DG&VJ).1	0.241	0.477
(AEHRC.A).1	0.199*	0.939
(WVU.AJ&VJ).1	0.142	0.448
(WVU.FP&VJ).1	0.112*	0.252
(UCDCSI.B.2)	0.006	0.035
<i>Task 1b, additional annotations:</i>		
(UCSC.CW&RA).2	0.545*	0.878
(UCSC.CW&RA).1	0.540*	0.879
(THCIB.A).1	0.470*	0.853
(WVU.AL&VJ).1	0.349*	0.625
(WVU.RK&VJ).1	0.247	0.531

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 more data and perform another evaluation in the near future.

## Acknowledgments

We greatly appreciate the hard work and feedback of our program committee members and annotators, including, but not limited to David Harris, Glenn Zaramba, Erika Siirala, Qing Zeng, Tyler Forbush, Jianwei Leng, Maricel Angel, Erikka Siirala, Heljä Lundgren-Laine, Jenni Lahdenmaa, Laura Maria Murtola, Marita Ritmala-Castren, Riitta Danielsson-Ojala, Saija Heikkinen, and Sini Koivula. This shared task was partially supported by Shared Annotated Resources (ShARe) project NIH 5R01GM090187, VAHSR&D HIR 08-374, NICTA (National Information and Communications Technology Australia), funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program), and NLM 5T15LM007059.

## References

1. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J.F., Leveling, J., Kelly, L., Goeriot, L., Martinez, D., Zuccon, G.: Overview of the share/clef ehealth evaluation lab 2013. In: Proceedings of ShARe/CLEF eHealth Evaluation Labs. (2013)
2. Mowery, D.L., South, B.R., Christensen, L., Murtola, L.M., Salanterä, S., Suominen, H., Martinez, D., Elhadad, N., Pradhan, S., Savova, G., , Chapman, W.W.: Task 2: Share/clef ehealth evaluation lab 2013. In: Proceedings of ShARe/CLEF eHealth Evaluation Labs. (2013)
3. Campbell, K., Oliver, D., Shortliffe, E.: The Unified Medical Language System: Towards a collaborative approach for solving terminologic problems. *J Am Med Inform Assoc* **5**(1) (1998) 12–16
4. Bodenreider, O., McCray, A.: Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics* **36** (2003) 414–432
5. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th Conference on Computational Linguistics (COLING), Saarbrücken, Germany (2000) 947–953