SemEval-2015 Task 14: Analysis of Clinical Text

Noémie Elhadad[♣], Sameer Pradhan[†], Sharon Lipsky Gorman[♣], Suresh Manandhar[◇], Wendy Chapman[♠], Guergana Savova[†]
[♣] Columbia University, USA
[†] Boston Children's Hospital, USA
[◇] University of York, UK
[♠] University of Utah, USA
noemie.elhadad@columbia.edu, guergana.savova@childrens.harvard.edu

Abstract

We describe two tasks—named entity recognition (Task 1) and template slot filling (Task 2)—for clinical texts. The tasks leverage annotations from the ShARe corpus, which consists of clinical notes with annotated mentions disorders, along with their normalization to a medical terminology and eight additional attributes. The purpose of these tasks was to identify advances in clinical named entity recognition and establish the state of the art in disorder template slot filling. Task 2 consisted of two subtasks: template slot filling given gold-standard disorder spans (Task 2a) and end-to-end disorder span identification together with template slot filling (Task 2b). For Task 1 (disorder span detection and normalization), 16 teams participated. The best system yielded a strict F1-score of 75.7, with a precision of 78.3 and recall of 73.2. For Task 2a (template slot filling given gold-standard disorder spans), six teams participated. The best system yielded a combined overall weighted accuracy for slot filling of 88.6. For Task 2b (disorder recognition and template slot filling), nine teams participated. The best system yielded a combined relaxed F (for span detection) and overall weighted accuracy of 80.8.

1 Introduction

Patient records are abundant with reports, narratives, discussions, and updates about patients. This unstructured part of the record is dense with mentions of clinical entities, such as conditions, anatomical sites, medications, and procedures. Identifying the different entities discussed in a patient record, their status towards the patient, and how they relate to each other is one of the core tasks of clinical natural language processing. Indeed, with robust systems to extract such mentions, along with their associated attributes in the text (e.g., presence of negation for a given entity mention), several high-level applications can be developed such as information extraction, question answering, and summarization.

In biomedicine, there are rich lexicons that can be leveraged for the task of named entity recognition and entity linking or normalization. The Unified Medical Language System (UMLS) represents over 130 lexicons/thesauri with terms from a variety of languages. The UMLS Metathesaurus integrates standard resources such as SNOMED-CT, ICD9, and RxNORM that are used worldwide in clinical care, public health, and epidemiology. In addition, the UMLS also provides a semantic network in which every concept in the Metathesaurus is represented by its Concept Unique Identifier (CUI) and is semantically typed (Bodenreider and McCray, 2003).

The SemEval-2015 Task 14, Analysis of Clinical Text is the newest iteration in a series of community challenges organized around the tasks of named entity recognition for clinical texts. In SemEval-2014 Task 7 (Pradhan et al., 2014) and previous challenge 2013 (Pradhan et al., 2013), we had focused on the task of named entity recognition for disorder mentions in clinical texs, along with normalization to UMLS CUIs. This year, we shift focus on the task of identifying a series of attributes describing a disorder mention. Like for previous challenges, we use

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 303–310,
Denver, Colorado, June 4-5, 2015.

the ShARe corpus[1] and introduce a new set of annotations for disorder attributes.

In the remainder of this paper, we describe the dataset and the annotations provided to the task participants, the subtasks comprising the overall task, and the results of the teams that participated along with notable approaches in their systems.

2 Dataset

	Train	Dev	Test
Notes	298	133	100
Words	182K	153K	109K

Table 1: Notes, words, and disorder distributions in the training, development, and testing sets.

The dataset used is the ShARe corpus (Pradhan et al., 2015). As a whole, it consists of 531 deidentified clinical notes (a mix of discharge summaries and radiology reports) selected from the MIMIC II clinical database version 2.5 (Saeed et al., 2002). Part of the ShARe corpus was released as part of Semeval 2014 Task 7. In fact, to enable meaningful comparisons of systems performance across years, the 2015 SemEval training set combines the 2014 training and development sets, while the 2015 SemEval development set consists of the 2014 test set. The 2015 test set is a previously unseen set of clinical notes from the ShARe corpus. Table 2 provides descriptive statistics about the different sets. In addition to the ShARe corpus annotations, task participants were provided with a large set of unlabeled deidentified clinical notes, also from MIMIC II (400,000+ notes).

The ShARe corpus contains gold-standard annotations of disorder mentions and a set of attributes, as described in Table 2. We refer to the nine attributes as a disorder template. The annotation schema for the template was derived from the established clinical element model[2]. The complete guidelines for the ShARe annotations are available on the ShARe website[3]. Here, we provide a few examples to illustrate what each attribute captures.

[1] share.healthnlp.org

[2] www.clinicalelement.com

[3] share.healthnlp.org

	Train	Dev
Disorder mentions	11,144	7,967
CUI=CUI-less	30%	24%
CUI	70%	76%
Unique CUIs	1,352	1,139
Negation = yes	19.6%	20.1%
Negation = no	80.4%	79.9%
Subject = patient	99.2%	98.4%
Subject = family_member	<1%	1.4%
Subject = other	<1%	<1%
Subject = donor_other	<1%	0%
Uncertainty = yes	8.9%	5.9%
Uncertainty = no	91.1%	94.1%
Course = changed	<1%	<1%
Course = resolved	<1%	<1%
Course = worsened	< 1%	<1%
Course = improved	< 1%	1%
Course = decreased	1.6%	<1%
Course = increased	2%	1.7%
Course = unmarked	94.1%	95.2%
Severity = slight	1.1%	<1%
Severity = severe	3.5%	2.6%
Severity = moderate	5.9%	2.3%
Severity = unmarked	89.49%	94.2%
Conditional = true	4.9%	6.2%
Conditional = false	95.1%	93.8%
Generic = true	<1%	1%
Generic = false	99.1	99%
Body Location = CUI	55.3%	44.7%
Body Location = null	44.4%	54.6%
Body Location = CUI-less	<1%	<1%
Unique BL CUIs	734	511

Table 3: Distribution of different attribute values in the training and testing sets.

- In the statement "patient denies numbness," the disorder numbness has an associated negation attribute set to "yes."
- In the sentence "son has schizophrenia", the disorder schizophrenia has a subject attribute set to "family_member."
- The sentence "Evaluation of MI." contains a disorder (MI) with the uncertainty attribute set to "yes".
- An example of disorder with a non-default course attribute can be found in the sentence "The cough got worse over the next two weeks.", where its value is "worsened."
- The severity attribute is set to "slight" in "He has slight bleeding."

Slot	Description	Possible Values
CUI	CUI; indicates normalized disorder	CUI, CUI-less
NEG	Negation; indicates whether disorder is negated	no*, yes
SUB	Subject; indicates who experiences the disorder	patient*, null, other, family_member, donor_family_member, donor_other
UNC	Uncertainty; indicates presence of doubt about the disorder	no*, yes
COU	Course; indicates progress or decline of the disorder	unmarked*, changed, increased, decreased, improved, worsened, resolved
SEV	Severity; indicates how severe the disorder is	unmarked*, slight, moderate, severe
CND	Conditional; indicates conditional existence of disorder under specific circumstances	false*, true
GEN	Generic; indicates a generic mention of a disorder	false*, true
BL	Body Location; represents normalized CUI of body location(s) associated with disorder	null*, CUI, CUI-less

Table 2: Disorder attributes and their possible values. Default values are indicated with an *.

- In the sentence "Pt should come back if any rash occurs," the disorder rash has a conditional attribute with value "true."
- In the sentence "Patient has a facial rash", the body location associated with the disorder "facial rash" is "face" with CUI C0015450. Note that the body location does not have to be a substring of the disorder mention, even though in this example it is.

The ShARe corpus was annotated following a rigorous process. Annotators were professional coders who trained for the specific task of ShARe annotations. The annotation process consisted of a double annotation step followed by an adjudication phase. For all annotations, in addition to all the values for the attributes, their corresponding character spans in the text were recorded and are available as part of the ShARe annotations. Table 3 shows the distribution of the different attributes in the training and development sets.

3 Tasks

The Analysis of Clinical Text Task is split into two tasks, one on named entity recognition, and one on template slot filling for the named entities. Participants were able to submit to either or both tasks.

3.1 Task 1: Disorder Identification

For task 1, disorder identification, the goal is to recognize the span of a disorder mention in input clinical text and to normalize the disorder to a unique CUI in the UMLS/SNOMED-CT terminology. The UMLS/SNOMED-CT terminology is defined as the set of CUIs in the UMLS, but restricted to concepts that are included in the SNOMED-CT terminology.

Participants were free to use any publicly available resources, such as UMLS, WordNet, and Wikipedia, as well as the large corpus of unannotated clinical notes.

The following are examples of input/output for Task 1.

1 In "The rhythm appears to be atrial fibrillation." the span "atrial fibrillation" is the gold-standard disorder, and its normalization is CUI C0004238 (preferred term atrial fibrillation). This is a
2 In "The left atrium is moderately dilated." the disorder span is discontiguous: "left atrium...dilated" and its normalization is CUI C0344720 (preferred term left atrial dilatation).
3 In "53 year old man s/p fall from ladder." the disorder is "fall from ladder" and is normalized to C0337212 (preferred term accidental fall from ladder).

Example 1 represents the easiest cases. Example 2 represents instances of disorders as listed in the UMLS that are best mapped to discontiguous mentions. In Example 3, one has to infer that the description is a synonym of the UMLS preferred term. Finally, in some cases, a disorder mention is present, but there is no good equivalent CUI in UMLS/SNOMED-CT. The disorder is then normalized to "CUI-less".

3.2 Task 2: Disorder Slot Filling

This task focuses on identifying the normalized value for the nine attributes described above: the CUI of the disorder (very much like in Task 1), negation indicator, subject, uncertainty indicator, course, severity, conditional, generic indicator, and body location.

We describe Task 2 as a slot-filling task: given a disorder mention (either provided by gold-standard or identified automatically) in a clinical note, identify the normalized value of the nine slots. Note that there are two aspects to slot filling: cues in the text and normalized value. In this task, we focus on normalized value and ignore cue detection.

To understand the state of the art for this new task, we considered two subtasks. In both cases, given a disorder span, participants are asked to identify the nine attributes related to the disorder. In Task 2a, the gold-standard disorder span(s) are provided as input. In Task 2b, no gold-standard information is provided; systems must recognize spans for disorder mentions and fill in the value of the nine attributes.

4 Evaluation Metrics

4.1 Task 1 Evaluation Metrics

Evaluation for Task 1 is reported according to a F-score, that captures both the disorder span recognition and the CUI normalization steps. We compute two versions of the F-score:

- *Strict F-score:* a predicted mention is considered a true positive if (i) the character span of the disorder is exactly the same as for the gold-standard mention; and (ii) the predicted CUI is correct. The predicted disorder is considered a false positive if the span is incorrect or the CUI is incorrect.
- *Relaxed F-score:* a predicted mention is a true positive if (i) there is any word overlap between the predicted mention span and the gold-standard span (both in the case of contiguous and discontiguous spans); and (ii) the predicted CUI is correct. The predicted mention is a false positive if the span shares no words with the gold-standard span or the CUI is incorrect.

Thus, given, D_{tp}, the number of true positives disorder mentions, D_{fp}, the number of false positive disorder mentions, and D_{fn}, the number of false negative disorder mentions

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

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

$$F = \frac{2 \times P \times R}{P + R} \quad (3)$$

4.2 Task 2 Evaluation Metrics

We introduce a variety of evaluation metrics, which capture different aspects of the task of disorder template slot filling. Overall, for Task 2a, we reported average unweighted accuracy, weighted accuracy, and per-slot weighted accuracy for each of the nine slots. For Task 2b, we report the same metrics, and in addition report relaxed F for span identification.

We now describe per-disorder evaluation metrics, and then describe the overall evaluation metrics which provide aggregated system assessment. Given the K slots $(s_1, ..., s_K)$ to fill (in our task the nine different slots), each slot s_k has n_k possible normalized values $(s_k^i) i \in 1..n_k$. For a given disorder, its gold-standard value for slot s_k is denoted gs_k, and its predicted value is denoted ps_k.

4.2.1 Per-Disorder Evaluation Metrics

Per-disorder unweighted accuracy The unweighted accuracy represents the ability of a system to identify all the slot values for a given disorder. The per-disorder unweighted accuracy is simply defined as:

$$\frac{\sum_{k=1}^{K} I(gs_k, ps_k)}{K}$$

where I is the identity function: $I(x, y) = 1$ if $x = y$ and 0 otherwise.

Per-disorder weighted accuracy The weighted per-disorder accuracy takes into account the prevalence of different values for each of the slots. This metric captures how good a system is at identifying rare values of different slots. The weights are thus defined as follows:

- The CUI slot's weight is set to 1, for all CUI values.
- The body location slot's weight is defined as weight(NULL) = 1-prevalence(NULL), and the weight for any non-NULL value (including CUI-less) is set to weight(CUI) = 1-prevalence(body location with a non-NULL value).

- For each other slot s_k, we define n_k weights $weight(s_k^i)$ (one for each of its possible normalized values) as follows:

$$\forall i \in 1..n_k, weight(s_k^i) = 1 - prevalence(s_k^i)$$

where $prevalence(s_k^i)$ is the prevalence of value s_k^i in the overall corpus(training, development, and testing sets). The weights are such that highly prevalent values have smaller weights and rare values have bigger weight.

Thus, weighted per-disorder accuracy is defined as

$$\frac{\sum_{k=1}^{K} weight(gs_k) * I(gs_k, ps_k)}{\sum_{k=1}^{K} weight(gs_k)} \quad (4)$$

where, like above, gs_k is the gold-standard value of slot s_k and ps_k is the predicted value of slot s_k, and I is the identity function: $I(x, y) = 1$ if $x = y$ and 0 otherwise.

4.2.2 Overall Evaluation Metrics

Weighted and Unweighted Accuracy. Armed with the per-disorder unweighted and weighted accuracy scores, we can compute an average across all true-positive disorders. For task 2a, the disorders are provided, so they are all true positive, but for task 2b, it is important to note that we only consider the true-positive disorders to compute the overall accuracy.

$$Accuracy = \frac{\sum_{i=1}^{\#tp} per\_disorder\_acc(tp_i)}{\#tp} \quad (5)$$

Per-Slot Accuracy. Per-slot accuracy are useful in assessing the ability of a system to fill in a particular slot. For each slot, an average per-slot accuracy is defined as the accuracy for each true-positive disorder to recognize the value for that particular slot across the true-positive spans. Thus, for slot s_k, the per-slot accuracy is:

$$\frac{\sum_{i=1}^{\#tp} weight(gs_{i,k}) * I(gs_{i,k}, ps_{i,k})}{\sum_{i=1}^{\#tp} weight(gs_{i,k})} \quad (6)$$

where for each true-positive span there is a gold-standard value $gs_{i,k}$ and a predicted value $ps_{i,k}$ for slot s_k.

team	run	strict_P	strict_R	strict_F	relax_P	relax_R	relax_F
ezDI	run 1	0.783	0.732	0.757	0.815	0.761	0.787
ULisboa	run 3	0.779	0.705	0.740	0.806	0.729	0.765
UTH-CCB	run 3	0.778	0.696	0.735	0.797	0.714	0.753
UWM	run 2	0.773	0.699	0.734	0.809	0.731	0.768
UTH-CCB	run 1	0.748	0.713	0.730	0.777	0.741	0.759
UTH-CCB	run 2	0.748	0.713	0.730	0.777	0.741	0.759
TAKELAB	run 1	0.761	0.696	0.727	0.794	0.727	0.759
ULisboa	run 2	0.749	0.681	0.713	0.780	0.709	0.743
Bioinformatics-UA	run 2	0.690	0.736	0.712	0.719	0.766	0.742
Bioinformatics-UA	run 3	0.691	0.735	0.712	0.720	0.765	0.742
ULisboa	run 1	0.748	0.676	0.710	0.782	0.706	0.742
CUAB	run 2	0.735	0.683	0.708	0.762	0.708	0.734
NYUClinicalML	run 3	0.741	0.676	0.707	0.775	0.707	0.740
Bioinformatics-UA	run 1	0.669	0.738	0.702	0.698	0.769	0.732
NYUClinicalML	run 1	0.722	0.662	0.691	0.763	0.699	0.729
NYUClinicalML	run 2	0.722	0.663	0.691	0.762	0.700	0.730
IHS-RD-Belarus	run 2	0.722	0.662	0.690	0.746	0.684	0.714
IHS-RD-Belarus	run 1	0.720	0.655	0.686	0.745	0.677	0.709
TeamHCMUS	run 1	0.680	0.633	0.656	0.711	0.662	0.685
TeamHCMUS	run 2	0.680	0.633	0.656	0.711	0.662	0.685
TeamHCMUS	run 3	0.680	0.633	0.656	0.711	0.662	0.685
CUAB	run 1	0.718	0.572	0.636	0.742	0.591	0.658
LIST-LUX	run 3	0.649	0.580	0.613	0.675	0.603	0.637
LIST-LUX	run 2	0.648	0.579	0.612	0.674	0.602	0.636
LIST-LUX	run 1	0.649	0.577	0.611	0.677	0.602	0.637
umlnlp2014	run 3	0.611	0.567	0.588	0.675	0.626	0.650
umlnlp2014	run 2	0.559	0.488	0.521	0.653	0.571	0.609
umlnlp2014	run 1	0.557	0.487	0.519	0.652	0.570	0.608
KPSCMI	run 1	0.429	0.565	0.488	0.472	0.620	0.536
UWM	run 1	0.760	0.258	0.385	0.813	0.276	0.412
TMUNSW	run 1	0.328	0.349	0.338	0.396	0.420	0.408
TMUNSW	run 2	0.321	0.340	0.330	0.387	0.410	0.398
TMUNSW	run 3	0.321	0.340	0.330	0.387	0.410	0.398
UtahPOET	run 2	0.295	0.315	0.305	0.352	0.376	0.364
UtahPOET	run 3	0.295	0.315	0.305	0.352	0.376	0.364
UtahPOET	run 1	0.270	0.306	0.287	0.344	0.390	0.366
Sanj-TUM	run 2	0.098	0.110	0.104	0.475	0.531	0.502
Sanj-TUM	run 3	0.098	0.110	0.104	0.444	0.496	0.469
Sanj-TUM	run 1	0.082	0.107	0.093	0.425	0.552	0.481

Figure 1: Task 1 results.

Disorder Span Identification. This overall metric is only meaningful for Task 2b, where the system has to identify disorders prior to filling in their templates. Like in Task 1, we report an F-score metric to assess how good the system is at identifying disorder span. Note that unlike in Task 1, this F score does not consider CUI normalization, as this is captured through the accuracy in the template filling task. Thus, a true disorder span is defined as any overalp with a gold-stand disorder span. In the case of several predicted spans that overlap with a gold-standard span, then only one of them is chosen to be true positive (the longest ones), and the other predicted spans are considered false positives.

5 Results

5.1 Task 1

16 teams participated in Task 1. Strict and relaxed precision, recall, and F metrics are reported in Figure 1. We relied on the strict F to rank different submissions. The best system from team ezDI reported

75.7 strict F, also reporting the highest relaxed F (78.7) (Pathak et al., 2015).

For disorder span recognition, most teams used a CRF-based approach. Features explored included traditional NER features: lexical (bag of words and bigrams, orthographic features), syntactic features derived from either part-of-speech and phrase chunking information or dependency parsing, and domain features (note type and section headers of clinical note). Lookup to dictionary (either UMLS or customized lexicon of disorders) was an essential feature for performance. To leverage further these lexicons, for instance, Xu and colleagues (Xu et al., 2015) implemented a vector-space model similarity computation to known disorders as an additional feature in their appraoch.

The best-performing teams made use of the large unannotated corpus of clinical notes provided in the challenge (Pathak et al., 2015; Leal et al., 2015; Xu et al., 2015). Teams explored the use of Brown clusters (Brown et al., 1992) and word embeddings (Collobert et al., 2011). Pathak and colleagues (Pathak et al., 2015) note that word2vec (Mikolov et al., 2013) did not yield satisfactory results. Instead, they report better results clustering sentences in the unannotated texts based on their sequence of part-of-speech tags, and using the clusters as feature in the CRF.

Teams continued to explore approaches for recognizing discontiguous entities. Pathak and colleagues (Pathak et al., 2015), for instance, built a specialized SVM-based classifier for that purpose.

For CUI normalization, the best performing teams focused on augmenting existing dictionaries with lists of unambiguous abbreviations (Leal et al., 2015) or by pre-processing UMLS and breaking down existing lexical variants to account for high paraphrasing power of disorder terms (Pathak et al., 2015).

5.2 Task 2

Six teams participated in Task 2a. Evaluation metrics are reported in Figure 2. We relied on the Weighted Accuracy (WA) to rank the teams (highlighted in the Figure is F*WA, but since in Task 2a gold-standard disorders are provided, F is 1). The best system (team UTH-CCB) yielded a WA of 88.6 (Xu et al., 2015).

For Task 2b, nine teams participated. Evaluation metrics are reported in Figure 3. We relied on the combination of F score for disorder span identification and Weighted Accuracy for template filling to rank the teams (F*WA in the figure). The best system (team UTH-CCB) yielded a F*WA of 80.8.

Approaches to template filling focused on building classifiers for each attribute. Specialized lexicons of trigger terms for each attribute (e.g., list of negation terms) along with distance to disorder spans was a helpful feature. Overall, like in Task 1, a range of feature types from lexical to syntactic proved useful in the template filling task.

The per-slot accuracies (columns BL, CUI, CND, COU, GEN, NEG, SEV, SUB, and UNC in Figures 2 and 3) indicate that overall some attributes are easier to recognize than others. Body Location, perhaps not surprisingly, was the most difficult after CUI normalization, in part because it also requires a normalization to an anatomical site.

6 Conclusion

In this task, we introduced a new version of the ShARe corpus, with annotations of disorders and a wide set of disorder attributes. The biggest improvements in the task of disorder recognition (both span identification and CUI normalization) come from leveraging large amounts of unannotated texts and using word embeddings as additional feature in the task. The detection of discontiguous disorder seems to still be an open challenge for the community, however.

The new task of template filling (identifying nine attributes for a given disorder) was met with enthusiasm by the participating teams. We introduced a variety of evaluation metrics to capture the different aspects of the task. Different approaches show that while some attributes are harder to identify than other, overall the best performing teams achieved excellent results.

Acknowledgments

This work was supported by the Shared Annotated Resources (ShARe) project NIH R01 GM090187. We greatly appreciate the hard work of our program committee members and the ShARe annotators.

Team	Run	F	A	F*A	WA	F*WA	BL	CUI	CND	COU	GEN	NEG	SEV	SUB	UNC
UTH-CCB	run 1	1.000	0.943	0.943	0.886	0.886	0.862	0.854	0.903	0.887	0.911	0.975	0.936	0.975	0.911
UTH-CCB	run 3	1.000	0.943	0.943	0.886	0.886	0.862	0.854	0.903	0.887	0.911	0.975	0.936	0.975	0.911
ezDI	run 1	1.000	0.934	0.934	0.880	0.880	0.812	0.918	0.695	0.887	0.887	0.916	0.803	0.960	0.854
UTH-CCB	run 2	1.000	0.953	0.953	0.876	0.876	0.862	0.854	0.817	0.811	0.873	0.975	0.899	0.964	0.834
UTU	run 3	1.000	0.945	0.945	0.857	0.857	0.825	0.827	0.823	0.798	0.888	0.970	0.915	0.920	0.853
UTU	run 2	1.000	0.944	0.944	0.855	0.855	0.814	0.827	0.823	0.798	0.888	0.970	0.915	0.920	0.853
UTU	run 1	1.000	0.939	0.939	0.846	0.846	0.775	0.827	0.822	0.792	0.888	0.964	0.918	0.923	0.857
UWM	run 2	1.000	0.859	0.859	0.818	0.818	0.531	0.911	0.838	0.802	0.836	0.924	0.895	0.933	0.831
TeamHCMUS	run 1	1.000	0.195	0.195	0.576	0.576	0.614	0.804	0.292	0.345	0.076	0.426	0.310	0.173	0.311
TeamHCMUS	run 2	1.000	0.195	0.195	0.576	0.576	0.614	0.804	0.292	0.345	0.076	0.426	0.310	0.173	0.311
TeamHCMUS	run 3	1.000	0.195	0.195	0.576	0.576	0.614	0.804	0.292	0.345	0.076	0.426	0.310	0.173	0.311
UtahPOET	run 3	0.936	0.795	0.744	0.476	0.446	0.457	0.234	0.483	0.814	0.838	0.845	0.759	0.908	0.660
UtahPOET	run 1	0.931	0.769	0.716	0.378	0.351	0.456	0.000	0.481	0.815	0.836	0.848	0.758	0.907	0.659
UtahPOET	run 2	0.931	0.769	0.716	0.378	0.351	0.456	0.000	0.481	0.815	0.836	0.848	0.758	0.907	0.659

Figure 2: Task 2a results.

References

Olivier Bodenreider and Alexa T McCray. 2003. Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, 36(6):414–432.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

André Leal, Bruno Martins, and Francisco Couto. 2015. ULisboa: Semeval 2015 - task 14 analysis of clinical text: Recognition and normalization of medical concepts. In *Proceedings of SemEval-2015*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Narayan Choudhary, and Amrish Patel. 2015. ezDI: A semi-supervised nlp system for clinical narrative analysis. In *Proceedings of SemEval-2015*.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy Chapman, and Guergana Savova. 2013. Task 1: Share/clef ehealth evaluation lab 2013. In *Online Working Notes of CLEF*, page 230.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62.

Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154.

Mohammed Saeed, C Lieu, G Raber, and RG Mark. 2002. Mimic II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in Cardiology, 2002*, pages 641–644. IEEE.

Jun Xu, Yaoyun Zhang, Jingqi Wang, Yonghui Wu, Min Jian, Ergin Soysal, and Hua Xu. 2015. UTH-CCB: The participation of the SemEval 2015 challenge - task 14. In *Proceedings of SemEval-2015*.

Team	run	F	A	F*A	WA	F*WA	BL	CUI	CND	COU	GEN	NEG	SEV	SUB	UNC
UTH-CCB	run 1	0.926	0.941	0.871	0.873	0.808	0.864	0.819	0.899	0.899	0.919	0.976	0.939	0.973	0.912
UTH-CCB	run 2	0.926	0.950	0.879	0.863	0.799	0.864	0.819	0.822	0.837	0.884	0.976	0.904	0.963	0.831
UTH-CCB	run 3	0.903	0.943	0.852	0.881	0.796	0.873	0.834	0.897	0.895	0.925	0.977	0.943	0.974	0.913
ezDI	run 1	0.915	0.935	0.856	0.868	0.795	0.826	0.858	0.816	0.866	0.921	0.978	0.812	0.911	0.857
UWM	run 2	0.893	0.852	0.761	0.798	0.713	0.532	0.858	0.839	0.794	0.845	0.932	0.907	0.929	0.838
CUAB	run 2	0.905	0.908	0.822	0.785	0.710	0.655	0.810	0.660	0.774	0.885	0.850	0.860	0.846	0.749
Bioinformatics-UA	run 2	0.853	0.884	0.754	0.814	0.695	0.691	0.866	0.697	0.856	0.889	0.807	0.877	0.819	0.800
Bioinformatics-UA	run 3	0.853	0.883	0.754	0.814	0.695	0.689	0.867	0.697	0.856	0.890	0.806	0.878	0.818	0.798
Bioinformatics-UA	run 1	0.843	0.883	0.745	0.813	0.686	0.692	0.864	0.697	0.857	0.887	0.807	0.878	0.811	0.799
TeamHCMUS	run 1	0.855	0.884	0.756	0.784	0.671	0.603	0.801	0.725	0.851	0.904	0.935	0.843	0.931	0.802
TeamHCMUS	run 2	0.855	0.884	0.756	0.784	0.671	0.603	0.801	0.725	0.851	0.904	0.935	0.843	0.931	0.802
TeamHCMUS	run 3	0.855	0.884	0.756	0.784	0.671	0.603	0.801	0.725	0.851	0.904	0.935	0.843	0.931	0.802
umlnlp2014	run 3	0.882	0.867	0.765	0.648	0.571	0.525	0.731	0.495	0.569	0.869	0.530	0.535	0.752	0.550
LIST-LUX	run 1	0.884	0.865	0.765	0.641	0.567	0.515	0.719	0.496	0.575	0.870	0.529	0.544	0.751	0.559
LIST-LUX	run 3	0.882	0.866	0.763	0.642	0.566	0.517	0.720	0.500	0.578	0.873	0.528	0.543	0.749	0.560
LIST-LUX	run 2	0.881	0.866	0.763	0.641	0.565	0.517	0.720	0.497	0.575	0.873	0.530	0.543	0.749	0.557
CUAB	run 1	0.839	0.873	0.732	0.669	0.561	0.523	0.784	0.490	0.564	0.855	0.543	0.522	0.736	0.539
umlnlp2014	run 2	0.820	0.864	0.708	0.641	0.526	0.511	0.732	0.482	0.547	0.882	0.516	0.521	0.761	0.544
umlnlp2014	run 1	0.820	0.864	0.708	0.640	0.525	0.511	0.730	0.482	0.547	0.882	0.516	0.521	0.761	0.544
UtahPOET	run 2	0.756	0.821	0.620	0.580	0.438	0.453	0.468	0.475	0.831	0.862	0.853	0.746	0.896	0.651
UtahPOET	run 3	0.756	0.821	0.620	0.580	0.438	0.453	0.468	0.475	0.831	0.862	0.853	0.746	0.896	0.651
UtahPOET	run 1	0.724	0.836	0.605	0.596	0.431	0.566	0.494	0.475	0.566	0.857	0.805	0.629	0.848	0.631
UWM	run 1	0.485	0.835	0.405	0.769	0.373	0.374	0.849	0.870	0.810	0.937	0.942	0.888	0.966	0.845

Figure 3: Task 2b results.