Evaluation Metrics for the SemEval-2015 Task 14: Analysis of Clinical Text

Noémie Elhadad, Sharon Lipsky Gorman, Sameer Pradhan, Guergana Savova,
Wendy Chapman, David Martinez, Danielle Mowery, Sumithra Velupillai, Lee Christiansen

November 4, 2014

The SemEval-2015 Analysis of Clinical Text is a follow up to the SemEval-2014 and ShARe-CLEF13, and ShARe-CLEF14 challenges. This year, there are two independent tasks: (1) identify a disorder mention in text; and (2) fill in slots corresponding to a disorder's attributes. This document describes the tasks and the evaluation metrics used to score participants' submissions.

1 Disorder Identification Task

In the disorder identification task, the goal is to recognize the span of a disorder mention and its normalization to the UMLS/SNOMED-CT terminology. Like in previous challenges, the evaluation consists of a general F-score, which represents both ability to do span recognition and normalization. There are two versions of the F-score.

- *Strict* F-score: a predicted mention is considered a true positive if (i) the span 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.

Given,

D_{tp} = Number of true positives disorder mentions,
D_{fp} = Number of false positive disorder mentions, and
D_{fn} = Number of false negative disorder mentions.

$$Precision = P = \frac{D_{tp}}{D_{tp} + D_{fp}} \tag{1}$$

$$Recall = R = \frac{D_{tp}}{D_{tp} + D_{fn}} \tag{2}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

2 Slot Filling Task

The slot filling task is as follows: Given a span of a disorder mention in a clinical note, identify the value for nine attributes of the disorder: CUI of the disorder, negation indicator, subject, uncertainty indicator, course, severity, conditional, generic indicator, and CUI(s) of body location. In this task, the span of the disorder might be given from the gold standard (Task 2a) or might be detected automatically like in the disorder identification task (Task 2b).

Note that there are two aspects to slot filling: cue and normalized value. In CLEF14, attention was given to both aspects (see `http://clefehealth2014.dcu.ie/task-2/2014-guidelines` for more details). In SemEval15, we focus on normalized value and ignore cue detection.

Attribute	Normalized value[1]
normalized disorder CUI	CUI, CUI-less (no default)
negation	no*, yes
subject	patient*, family_member, donor_family_member, donor_other, null, other
uncertainty	no*, yes
course	unmarked*, changed, increased, decreased, improved, worsened, resolved
severity	unmarked*, slight, moderate, severe
conditional	false*, true
generic	false*, true
body location	NULL*, CUI, CUI-less

Next we describe the *p*er-disorder evaluation metric for this task. Given K slots $(s_1, ..., s_K)$, each slot s_k has n_k possible normalized values $(s_k^i)_{i \in 1..n_k}$, including its the default value s_k^d.

For a given disorder, the gold-standard value of one of its slot s_k is denoted gs_k, and its predicted value is denoted ps_k.

Example 1. Assume a task with 5 slots. A given concept has the following gold-standard and predicted slots.

Slot s_k	GS value gs_k	Pred value ps_k	Prevalence of GS value in corpus
Negation	negated	negated	10%
Subject	family	patient	5%
Uncertainty	unspecified (default)	uncertain	70%
Generic	specific (default)	specific	95%
Conditional	non-conditional (default)	non-conditional	97%

2.1 Per-Disorder Evaluation Metrics

2.1.1 Per-Disorder Unweighted Accuracy

The unweighted accuracy represents the accuracy of slot filling for all slots in the gold-standard annotation. For a given concept, this metric is defined as:

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

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

Unweighted relaxed accuracy for Example 1. For this concept, the metric will be computed as: $(1 + 0 + 0 + 1 + 1)/5 = 0.6$

2.1.2 Per-Disorder Weighted Accuracy

The weighted accuracy takes into account the prevalence of different values for each slots.

- In the case of normalized CUI, the weight is set to 1
- In the case of body location, the weight is set to weight(NULL) = 1-prevalence(NULL) in the entire corpus, 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:

$$\forall i \in 1..n_k, weight(s_k^i) \equiv 1 - prevalence(s_k^i) \tag{5}$$

  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 lower weights and rare values have higher weight.

For a given disorder, the weighted accuracy is defined as:

$$\frac{\sum_{k=1}^{K} weight(gs_k) * I(gs_k, ps_k)}{\sum_{k=1}^{K} weight(gs_k)} \tag{6}$$

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.

Weighted accuracy for Example 1. Given the prevalence listed in Example 1, the weights for the gold standard slots are:
$weight(Negation, negated) = 0.9$
$weight(Subject, family) = 0.95$
$weight(Uncertainty, unspecified^*) = 0.3$
$weight(Generic, specific^*) = 0.05$
$weight(Conditional, non-conditional^*) = 0.03$

and the weighted accuracy for the concept is:

$$\frac{0.9 * 1 + 0.95 * 0 + 0.3 * 0 + 0.05 * 1 + 0.03 * 1}{0.9 + 0.95 + 0.3 + 0.05 + 0.03} = 0.4394 \tag{7}$$

2.2 Overall Evaluation Metrics

There are two evaluation setups for the slot filling task:

Task 2a Provide the gold-standard information for disorders (span), and the participant has to fill the slots (including the CUI of the disorder).

Task 2b End-to-end: No gold-standard information is provided, and the participant has to (i) identify disorders (i.e., span recognition), and (ii) fill the slots for the disorder (including normalized disorder). As such, all identified disorders are counted in the evaluation.

In both cases, the evaluation of the normalization of the disorder to a CUI is taken care of as part of the slot filling task.

The following metrics are computed:

F-measure for span identification. In Task 2a, since the gold-standard disorder spans are provided, the F-measure will be always 1. The following describes the comptuation for Task 2b. A true positive disorder span is defined as any overlap with a gold-standard span – this is the same definition as for the Relaxed F for the Span Identifiction Task. If there are several predicted spans that overlap with a gold-standard span, then only one of them is chosen to be true positive (the longest span), and the other predicted spans are considered false positives.

$$P = \frac{\#tp}{\#tp + \#fp} \tag{8}$$

$$R = \frac{\#tp}{(\#tp + \#fn)} \tag{9}$$

$$F = \frac{2 * P * R}{P + R} \tag{10}$$

Weighted and Unweighted Accuracies. For boths subtasks, the weighted accuracy is defined as the average per-disorder accuracy (either weighted or unweighted) across the true-positive spans.

$$Accuracy = \frac{\sum_{i=1}^{\#tp} per\_disorder\_acc(tp_i)}{\#tp} \tag{11}$$

Per-Slot Accuracy. 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. For slot k, the per-slot accuracy is thus:

$$Per\_slot\_Accuracy_k = \frac{\sum_{i=1}^{\#tp} weight(gs_{i,k}) * I(gs_{i,k}, ps_{i,k})}{\sum_{i=1}^{\#tp} weight(gs_{i,k})} \tag{12}$$

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

These per-slot accuracies are useful in assessing the ability of a system to fill in a particular slot.

Task 2b Combined evaluation metric. For the end-to-end setup, it is useful to combine the F-measure computed for span identification and the Accuracy for the slot filling task.

$$F * WeightedAccuracy; F * UnweightedAccuracy \tag{13}$$

Example for Task 2b. Assume there are 5 gold-standard disorders in the test corpus. The system predicts 3 true positive disorder spans (tp), 5 false positive spans (fp), and 2 false negative spans (fn). Of note, for the definition of true positive, please refer to previous page.

The precision, recall, and F-measure for the span identification is therefore:

$$P = \frac{\#tp}{\#tp + \#fp} = \frac{3}{(3+5)} = 0.375 \tag{14}$$

$$R = \frac{\#tp}{(\#tp + \#fn)} = \frac{3}{3+2} = 0.6 \tag{15}$$

$$F = \frac{2 * P * R}{P + R} = 0.46 \tag{16}$$

Assume that for the three true-positive disorders, the system obtained the following per-disorder accuracy: 0.3, 0.8, and 0.7. Then, the slot-filling accuracy is:

$$Acc = \frac{0.3 + 0.8 + 0.7}{3} = 0.6 \tag{17}$$

and the overall metric for the task is:

$$F * Acc = 0.46 * 0.6 = 0.276 \tag{18}$$