Evaluation Data, Metrics and Software Resources

Samuel Bayer
The MITRE Corporation

1 Summary

The Adverse Drug Event Evaluation (ADE Eval) is an evaluation of tools to identify adverse events (AEs) mentioned in publicly available drug labels. The FDA CDER Office of Surveillance and Epidemiology (OSE) is sponsoring this ADE Eval. OSE is interested in a tool that would enable pharmacovigilance safety evaluators to automate the identification of labeled AEs which could facilitate triage, review and processing of safety case reports.

This evaluation uses a definition of adverse drug events specific to the business process in the Office of Surveillance and Epidemiology. The task will consist of identifying OSE-defined adverse drug events and mapping them to associated terms in the Medical Dictionary for Regulatory Activities (https://www.meddra.org) for specific sections of drug labels.

This document describes the evaluation data, metrics and software resources related to the ADE Eval evaluation of adverse event annotation and coding. The evaluation metrics address two separate use cases of interest to OSE and cover both mention-level extraction of AEs and their MedDRA encoding. The training data for this evaluation includes both AEs which should be found (labeled "OSE_Labeled_AE") and AEs which might be confusable with those AEs.

2 Data

The data for the evaluation consists of

- 100 annotated training documents, in an XML schema described below.
- 2000 unannotated test documents, in the same XML schema. 100 of these have been annotated for evaluation, but the identity of this 100-document subset will not be revealed, and these documents will not be released after the evaluation.

Each document consists of a subset of the sections found in a single drug label from DailyMed (https://dailymed.nlm.nih.gov/dailymed/). The sections of interest are adverse event, boxed warnings, and either one or two sections devoted to warnings and precautions. All documents contain an adverse event section; the other sections may or may not appear in a given document. The sections have been extracted from the DailyMed XML and converted to raw text using utilities developed originally for the NIST TAC 2017 Adverse Drug Reaction Extraction from Drug Labels evaluation (https://bionlp.nlm.nih.gov/tac2017adversereactions/). 50 of the training documents are shared with the TAC 2017 test set, although they have been reannotated to conform to the guidelines for the current evaluation.

3 Data format

The data format for the gold-standard documents and the expected submission documents differ slightly. We will not provide a DTD for these formats, but we will provide a wellformedness-checking tool, as well as a Python API for creating the submissions for those who are interested. Where possible, the data format has been preserved from the NIST TAC 2017 evaluation.

3.1 Gold standard format

We describe the gold-standard document format first. We exemplify using portions of the ANORO drug label, from the training set.

```
<?xml version="1.0" encoding="UTF-8"?>
<GoldLabel drug="ANORO">
  <Text>
    <Section id="S1" name="adverse reactions">...</Section>
    ...
  </Text>
  <IgnoredRegions>
    <IgnoredRegion len="19" name="heading" section="S1" start="4" />
    ...
  </IgnoredRegions>
  <Mentions>
    <Mention id="M49" len="6" reason="class_effect" section="S1"
             start="122" type="OSE_Labeled_AE">
      <Normalization meddra_llt="Asthma aggravated" meddra_llt_id="10003554"
                     meddra_pt="Asthma" meddra_pt_id="10003553" />
    </Mention>
    ...
    <Mention id="M5" len="15,8" reason="general_term" section="S1"
             start="4798,4836" type="NonOSE_AE">
      <Normalization meddra_llt="Musculoskeletal disorder"
                     meddra_llt_id="10048592"
                     meddra_pt="Musculoskeletal disorder"
                     meddra_pt_id="10048592" />
    </Mention>
    ...
  </Mentions>
</GoldLabel>
```

The toplevel element name is GoldLabel, and it contains three children: Text, IgnoredRegions, and Mentions. Text, in turn, contains up to four Section elements. Each Section has an ID and a name. The possible names are “adverse reactions”, “boxed warnings”, “warnings and precautions”, “warnings”, “precautions”. These names correspond to the section names in the original DailyMed drug label.

The IgnoredRegions section contains zero or more IgnoredRegion elements, each of which indicates a region of text in which all annotations will be ignored. The adverse event mentions in these regions are largely redundant and are generally of no interest to OSE. Performers will

be neither rewarded or penalized for producing mentions in these regions. The start and len attributes are character, not byte, counts. The two possible region names are “heading” and “excerpt”. The value of the section attribute is the ID of the corresponding section. In the example above, the ignored region begins at character 4 of section S1 and is 19 characters long.

The Mentions section contains zero or more Mention elements, each of which contains one or more Normalization elements. These are the elements that performers will provide. The attribute values of the Mention element are:

- id: the ID of the element. This ID must start with “M”.
- section: the ID of the section in which the mention appears.
- start: the character offset of the beginning of the mention. If the mention is discontinuous, the starts of the individual subspans will be comma-separated, as exemplified above.
- len: the length of the mention, in characters. If the mention is discontinuous, the lengths of the individual subspans will be comma-separated, as exemplified above.
- type: one of “OSE_Labeled_AE” (adverse events of interest to OSE), “NonOSE_AE” (adverse events or similar textual elements that are not of interest to OSE), “Not_AE_Candidate” (spans that may appear to be adverse events but are not). The latter two types are provided to performers for reference; performers will not be asked to produce them, and any mentions found in submission with a type other than "OSE_Labeled_AE" will be discarded.
- reason: the reason for the type assignment. These reasons are provided to the performers for reference; performers will not be asked to produce them, and any reason attributes found in submissions will be discarded. The values for this attribute are described below.

The Mention element may contain one or more Normalization elements. When multiple Normalization elements are present, they are judged to be equally valid MedDRA encodings for the mention, and a submission Normalization only needs to match one of them. The attributes of the Normalization element are:

- meddra_llt: the MedDRA low-level term which best matches the mention
- meddra_llt_id: the ID of that low-level term
- meddra_pt: the MedDRA preferred term which best matches the mention
- meddra_pt_id: the ID of that preferred term

3.1.1 Normalization attribute guidance

The version of MedDRA used during gold standard annotation is 20.1 (English). This is the version of MedDRA you should use for your submissions.

Each gold standard Mention element will contain at least one Normalization element which has values for all four attributes, and all gold standard Normalization elements will have values for both the meddra_pt and meddra_pt_id attributes. However, in some cases, additional Normalization elements within a Mention will have values only for meddra_pt and meddra_pt_id.

Of the attributes in the Normalization element, only the meddra_pt_id attribute will be used for scoring. However, you are required to provide values for both the meddra_pt and meddra_pt_id attributes, for the purpose of analysis and wellformedness checking.

3.2 Mention reasons

The values for the reason attribute in Mention elements are described in this table.

Type	Reason	Description
OSE_Labeled_AE	from_drug_use	AE associated with the use of the drug of interest
OSE_Labeled_AE	from_drug_component	AE associated with an inactive ingredient of the drug of interest.
OSE_Labeled_AE	class_effect	Effect associated with a drug class to which the drug of interest belongs.
OSE_Labeled_AE	medication_error	Preventable event that may cause or lead to inappropriate medication use or patient harm while the medication is in the control of a healthcare provider, patient, or consumer. Examples include administration of incorrect dosage, taking a drug at incorrect frequency, or administering the wrong drug because if easily confused names.
OSE_Labeled_AE	other	A reason for interest other than the specific reasons listed.

NonOSE_AE	manifestation_or_complication	Text describing signs, symptoms, or changes in lab results related to the manifestations of an AE and the sequelae of an AE are not of interest.
NonOSE_AE	AE_rate_lteq_placebo	AEs with incidence rate equal to or lower than placebo are not of interest.
NonOSE_AE	AE_animal	AEs observed in animal data are not of interest.
NonOSE_AE	AE_from_drug_interaction	AEs that result from drug-drug interaction or co-administration are not of interest.
NonOSE_AE	general_term	General terms or non-specific text such as broad categories (e.g. MedDRA system organ class) used to introduce AEs or text describe an outcome (e.g., death) rather than an AE. These are not of interest.
NonOSE_AE	AE_from_off_label	AEs associated with off-label or unapproved drug use are not of interest.
NonOSE_AE	AE_only_as_instruction	AEs mentioned in instructions are often mentioned in a hypothetical context, with instructions for or what to do if they develop. These AEs are not of interest.
NonOSE_AE	AE_for_another_drug_in_class	AE is related to a class effect but the label specifically states that

		the AE has not been reported or associated with the drug of interest or that the AE was only reported for another drug in the class to which the drug of interest belongs. These AEs are not of interest.
NonOSE_AE	OD_or_withdrawal	AE associated with discontinuing a medication or taking more than the prescribed amount. Drug overdose or withdrawal do not generally occur when a drug is used as indicated. Additionally, in the context of pharmacovigilance, identifying AEs associated with the drug when used as indicated is the highest priority. These AEs are not of interest.
NonOSE_AE	negation	AE whose presence or occurrence is negated or denied. These AEs are not of interest.
NonOSE_AE	other	A reason for disinterest other than the specific reasons listed.
Not_AE_Candidate	indication	A clinical symptom or circumstance for which the use of the drug of interest would be appropriate. These mentions are not AEs.

Not_AE_Candidate	contraindication	A clinical symptom, circumstance, or condition for which the use of the drug of interest would be inappropriate. These mentions are not AEs.
Not_AE_Candidate	preexisting_condition_or_risk_factor	Mentions that describe a condition that developed prior to applying the medication of interest, or condition that increases the likelihood of developing a disease or injury. These mentions are not AEs.
Not_AE_Candidate	other	A reason other than the specific reasons listed that the mention is not actually an AE.

3.3 Submission format

The submission format differs from the gold-standard format in the following ways:

- The toplevel element is SubmissionLabel rather than GoldLabel.
- The only recognized value for the type attribute of the Mention element is "OSE_Labeled_AE". Mentions with other type values will be discarded.
- The reason attribute of the Mention element will be discarded.
- The Mention element should contain exactly one Normalization element; if multiple Normalization elements are provided, all values other than the first value will be discarded.

The 2000 labels in the test set will contain all relevant Section and IgnoredRegion elements; performers will not be asked to discover either of these elements. Each label will contain an empty Mentions element which must be populated. The performers must respect and preserve the section IDs provided in these labels.

4 Metrics

The evaluation envisions two distinct use cases. The metrics assigned with all the use cases require the submission and gold mentions first to be paired with each other; the scorer will use a standard set alignment algorithm (the Kuhn-Munkres, or Hungarian, algorithm) to compute

the best pairing based on a mention similarity measure which is 80% overlap percentage and 20% MedDRA code match.

4.1 "Back office" use case

In the back office use case, the idea is that if OSE uses an automated system to identify AEs, some of the drug labels which are annotated by the system would be hand-corrected by human annotators for further improvement of the system. In this use case, every mention is important, and some corrections are more time-consuming than others.

Therefore, the primary metric for this use case is weighted, micro-averaged, corpus-level recall/precision/f-measure on mentions. In this use case, a perfect score will be awarded to a submission mention which is paired with a gold mention which matches exactly in extent, and bears the proper MedDRA code. In addition, partial matches and other errors will be weighted according to a generalization such as the following:

- deleting a spurious mention is easy (current assigned weight is .25)
- correcting a span issue or a MedDRA code is harder (current assigned weight is .5)
- adding a missing mention is harder still (current assigned weight is 1)

4.2 "Front office" use case

In the front office use case, safety evaluators need to know whether a MedDRA code is present in a given section and may want to see evidence of the presence of this MedDRA code. The section-level analysis is crucial because the SE needs to know what level of severity the drug label reflects for the given AE, and the sections differ in severity.

For this use case, we compute two primary metrics.

The first metric is macro-averaged recall/precision/f-measure on MedDRA codes. The scope of the macro-averaging is the section; R/P/F will be computed per section and the values will be averaged together.

In this computation, a correct MedDRA code is one which is realized in the gold standard by at least one mention which is paired with a submission mention with the same code. (We will refer to these codes as "properly grounded".) The two mentions do not need to match in extent. All other MedDRA codes are judged to be either missing (i.e., it is the MedDRA code for a mention in the gold standard, but no mention pair that is coded in this way contains a submission mention with a matching code) or spurious (i.e., it is the MedDRA code for a mention in the submission, but no mention pair that is coded in this way contains a gold standard mention with a matching code).

The second metric is a "quality" metric on the properly grounded MedDRA codes from the first metric. This metric is a macro-averaged precision measure on submission mentions associated with each correct code. Each correct MedDRA code is assigned a score of the percentage of the

submission mentions linked to it which are paired with an identically-coded gold mention, multiplied by the percentage overlap of each mention with its gold pair. The score for the MedDRA codes within a section are averaged to create the section-level score, and these scores are macro-averaged across the corpus.

To illustrate, we provide the following example. Assume the submission section contains 9 mentions, divided as follows:

- Mentions 1 - 5 are coded to MedDRA code A, which is present in the gold. Mention 2 overlaps with a gold mention which is coded to MedDRA code A; mentions 1, 3, 4, 5 do not overlap with any gold mention.
- Mentions 6 and 7 are coded to MedDRA code B, which is present in the gold. Mention 6 overlaps with a gold mention which is coded to MedDRA code C. Mention 7 overlaps with a gold mention which is coded to MedDRA code B.
- Mentions 8 and 9 are coded to MedDRA code D, which is not present in the gold. Mention 8 overlaps with a gold mention which is coded to MedDRA code B. Mention 9 does not overlap with any gold mention.

At issue are four MedDRA codes. A, B, and D are present in the submission, and A, B and C are present in the gold.

Of the MedDRA codes, MedDRA code A is properly grounded, because of submission mention 2. MedDRA code B is properly grounded, because of submission mention 7. MedDRA code D is not properly grounded, because there are no mentions of MedDRA code D in the gold. MedDRA code C is not properly grounded, because no mentions in the submission are coded to MedDRA code C.

So for the purposes of the first "front office metric", this section contains 2 matching codes, 1 spurious code (MedDRA code D), and one missing code (MedDRA code C). Recall, precision, and f-measure are then computed as usual.

For the purposes of the quality metric, we compute the overlap weight as the size of the intersection of the character offsets covered by the submission and gold divided by the size of the union of the character offsets. The quality metric is computed for each of the properly grounded MedDRA codes A, B. The score for MedDRA code A, if the span of mention 2 matches its gold pair exactly, would be .2 (1 of 5 mentions is correct), less if the span match is not exact. The score for MedDRA code B, if the span of mention 7 matches its gold pair exactly, would be .5 (1 of 2 mentions is correct), less if the span match is not exact. The average of these individual code scores is the quality score of the section.

5 Submission

Each performer team may submit up to two submissions. Preferably, each submission should be a zip file of a single folder containing the 2000 test labels. The submissions must be

accompanied by a two-page system description that includes descriptions of the architecture (including an architecture diagram), and the data and other resources used in developing the system.

6 Software resources

Performers will be provided with the following software resources:

- a wellformedness checker to use to check their submissions. Any submissions which do not pass the wellformedness checker will be rejected; no exceptions. Be sure to run the checker before you submit.
- a scorer which implements the metrics described above.
- a Python 2 API which performers can use to construct their submissions, if they choose
- a tool which translates the evaluation XML into standalone annotation visualizations using the brat annotation tool (http://brat.nlplab.org/index.html).

Documentation for these resources will be provided.