i2b2 Medication Challenge
(Tentative) Evaluation Metrics
16 June 2009
Ozlem Uzuner, Imre Solti, Fei Xia

The evaluation metrics of the i2b2 Medication Challenge are adapted from the evaluations of the Question Answering Track of TREC. Evaluation scripts will be released before the end of the development period. All evaluations assume correctly formatted output, e.g., fields correctly labeled and offsets properly separated from the extracted text in each field.

We use two kinds of evaluation metrics:
1. Strict evaluation with exact matches
2. Relaxed evaluation with inexact matches

1. Exact Evaluation: For each list, compute instance precision, instance recall, and F-measure as described in http://trec.nist.gov/pubs/trec12/papers/QA.OVERVIEW.pdf.

The instance precision (IP) and instance recall (IR) for a list can be computed as follows: Let S be the size of the ground truth list (i.e., the number of known instances), D be the number of correct, distinct instances returned by the system as determined by exact offset matching, and N be the total number of instances returned by the system. Then:

$$IP = \frac{D}{N}$$

$$IR = \frac{D}{S}$$

$$F = \frac{(\beta^2 + 1) * IP * IR}{(\beta^2 * IP) + IR} \text{ where } \beta = 1$$

2. Inexact Evaluation: Given a list, inexact recall is the proportion of system-returned tokens that overlap with the ground truth tokens.

Inexact Recall = # correctly returned tokens from each instance as determined by inexact matching / # tokens in the ground truth

Inexact precision is length based. Given the length of the list of medications (token count), inexact precision for the list is:

Inexact Precision = # tokens from each instance in system output that match ground truth / # tokens in system output

The inexact evaluation will rely on the F-measure formula for exact evaluation, but use inexact precision instead of instance precision and inexact recall instead of instance recall.

Consider the following sample output:

```
m="caltrate plus d" 5:1 5:3 || do="one" 5:4 5:4 || mo="p.o" 5:6 5:6 ||
f="b.i.d." 5:7 5:7 || du="nm"||r="nm"||…
m="lantus 7" 6:1 6:2||do="units" 6:2 6:3 || mo="nm"|| f="q.p.m" 6:5 6:5
|| du="nm" || r="nm"||…
m="novolog" 7:1 7:1||do="4 units/4 units/5 units" 7:2 7:5 || mo="sc"
7:6 7:6 ||f="t.i.d." 7:7 7:7 ||du="nm"||r="nm"||…
```

Assume that the matching ground truth is as follows:

```
m="caltrate plus d" 5:1 5:3 || do="one tab" 5:4 5:5 || mo="nm" ||
f="b.i.d." 5:7 5:7 || du="nm"||r="nm"||…
m="lantus" 6:1 6:1||do="7 units" 6:2 6:3 || mo="sc" 6:4 6:4 ||
f="q.p.m" 6:5 6:5 || du="nm" || r="nm"||…
m="novolog" 7:1 7:1||do="4 units/4 units/5 units" 7:2 7:5||mo="sc" 7:6
7:6 ||f="t.i.d." 7:7 7:7 ||du="nm"||r="nm"||…
m="imdur" 8:1 8:1||do="30 mg" 8:2 8:3||mo="nm"||f="b.i.d." 8:4
8:4||du="nm"|| …
```

Entries that are `nm` in the ground truth are omitted from the evaluation. We eliminate duplicate entries from the evaluation. In order to be considered duplicate, two entries need to be exactly the same in all of their fields.

We align entries in the system output with entries in the ground truth using the medication name and offset. In case of unique entries (one in system output and one in ground truth) for each medication at a given offset, matching on the medication name and offset creates a one-to-one correspondence between the ground truth entries and the entries in the system output. In case of multiple entries for a medication at a given offset, we use a greedy approach to align the entries in the system output with entries in the ground truth. For each entry of the system output, the ground truth entry that gives the best F-measure is selected as the alignment match; each ground truth entry can match only one entry in the system output. In case of inexact matches of medication names and offsets, greedy approach and F-measure is used to find the best matching ground truth entry for each entry in the system output; each ground truth entry is matched to only one entry in the system output.

Given aligned system outputs, we perform two kinds of evaluation:

1. Horizontal
2. Vertical

We perform evaluations at two different levels of granularity:

1. Patient record level
2. System level

System level horizontal evaluation is the primary evaluation metric. Two sets of rankings will be provided; one for exact and one for inexact evaluation.

1. Horizontal evaluation:

Consider list 1 in system output above:

```
m="caltrate plus d" 5:1 5:3
do="one" 5:4 5:4
mo="p.o" 5:6 5:6
f="b.i.d." 5:7 5:7

… <rest of the fields are "nm">
```

Corresponding entry in the ground truth:

```
m="caltrate plus d" 5:1 5:3
do="one tab" 5:4 5:5
f="b.i.d." 5:7 5:7
… <rest of the entries are "nm">
```

i. Exact evaluation:
   e.g., for list 1 and its ground truth above:
   N=4 (total number of fields in the system output that are not `nm`)
   D=2 (2 exact matches in terms of offsets and field type)
   S=3 (total number of fields in the ground truth that are not `nm`)
   Instance Precision = 2/4
   Instance Recall = 2/3

ii. Inexact evaluation: For the same two lists, the inexact evaluation gives:
   Inexact Recall = 5/6
   Inexact Precision = 5/6

a. Patient record level evaluation:
   i. Exact evaluation: Micro-average over all entries in a single record, taking into consideration the link between fields that belong to a single entry. Macro-average over all records.
   ii. Inexact evaluation: Micro-average over all entries in a single record, taking into consideration the link between fields that belong to a single entry. Macro-average over all records.
b. System level evaluation (Primary evaluation metrics):
   i. Exact evaluation: Micro-average over all entries in the system output, taking into consideration the link between the fields that belong to a single entry and the entries that belong to a single record.
   ii. Inexact evaluation: Micro-average over all entries in the system output, taking into consideration the link between the fields that belong to a single entry and the entries that belong to a single record.

2. Vertical Evaluation:
We create separate lists for each type of field and remove duplicates from each list. e.g., for the above sample output, we have the following lists:

```
medications:
m="caltrate plus d" 5:1 5:3
m="lantus 7" 6:1 6:2
m="novolog" 7:1 7:1

dosages:
do="one" 5:4 5:4
do="units" 6:3 6:3
do="4 units/4 units/5 units" 7:2 7:5

modes:
mo="p.o" 5:6 5:6
mo="sc" 7:6 7:6

frequencies:
f="b.i.d." 5:7 5:7
f="q.p.m" 6:5 6:5
f="t.i.d." 7:7 7:7
```

The ground truth above gets converted to:

```
medications:
m="caltrate plus d" 5:1 5:3
m="lantus" 6:1 6:1
m="novolog" 7:1 7:1
m="imdur" 8:1 8:1

dosages:
do="one tab" 5:4 5:5
do="7 units" 6:2 6:3
do="4 units/4 units/5 units" 7:2 7:5
do="30 mg" 8:2 8:3

modes:
mo="sc" 6:4 6:4
mo="sc" 7:6 7:6

frequencies:
f="b.i.d." 5:7 5:7
f="q.p.m" 6:5 6:5
f="t.i.d." 7:7 7:7
f="b.i.d." 8:4 8:4
```

i. Exact Evaluation: Evaluate each list separately.

Instance precision for medications = 2/3
Instance precision for dosages = 1/3
Instance precision for modes = 1/2
…
Instance recall for medications = 2/4

Instance recall for dosages = 1/4
Instance recall for modes = 1/2
…

ii. Inexact: Same formulation as exact evaluation but with the metrics for inexact evaluation.

Inexact precision for medications = 5/6
Inexact precision for dosages = 6/6 = 1
Inexact precision for modes = 1/2
…
Inexact recall for medications = 5/6
Inexact recall for dosages = 6/10
Inexact recall for modes = 1/2
…

a. Patient record level evaluation:
Create one list per field type per patient record.

i. Exact: Micro-average over all field types in the (ground truth) record. Macro-average over all records in the ground truth.
ii. Inexact: Micro-average over all field types in the (ground truth) record. Macro-average over all records in the ground truth.

b. System level evaluation:
Create one list per field type per system.

i. Exact: Micro-average over all field types in ground truth.
ii. Inexact: Micro-average over all field types in ground truth.

List vs. Narrative distinction:
While system ranking will be on the complete system output, we will also apply the above evaluation metrics to list and narrative entries separately, in order to get a sense of how well the systems did on the examples extracted from narratives.

Formulae for micro- and macro-averaged F-measure:

M is the number of records.

Equation 1 – Macro-averaged F-measure (F_{1macro}) $$F_{1_{macro}} = \frac{\sum_{i=1}^{M} F_{1_i}}{M}$$

Micro–averaged F-measure over all entries in a record is:

Equation 2 – Micro-averaged F-measure (F_{1micro}) $$F_{1_{micro}} = \frac{(\beta^2 + 1) * IP * IR}{(\beta^2 * IP) + IR}$$ where $\beta = 1$

The formula above can be adapted for calculating the micro-averaged F-measure over all entries in the system output.

One uncertain point:

- We are hoping that we (the organizers) will be able to create a small set of ground truth documents that mimic the output in the released sample annotations. This will be evaluation set A.

- Majority of the annotations on the test data will be provided by the challenge participants. You can expect that each record will be annotated by three independent teams. The disagreements will be resolved by the organizers.

    - Some of the disagreements will be resolved manually. The result of this will be evaluation set B.
    - In case of shortage of time and resources of the organizers, an automatic approach may be employed for determining the reference set from the participant annotations. Full details will be exposed as they are determined. The result of this will be evaluation set C.

We expect that evaluation set A will be the smallest in size but will have the highest confidence. Evaluation set B will be larger in size than evaluation set A, but will have lower confidence. Evaluation set C will most likely be the largest set but will have the lowest confidence.

Performance against each evaluation set will be measured separately (as described in the earlier pages). For system ranking purposes we will combine the results of the evaluation sets for each system, in addition to reporting results on each of these sets.

Currently, simple arithmetic average of the performances on the three evaluation sets is the strongest candidate for final evaluation of each system. We admit that this metric is biased towards confidence rather than evaluation set size.