**BANNER-improvement project 03/28/2014**

*Proposal*

Table of Contents

I. Objective 1

II. Data 1

III. Contest options and suggestions 2

IV. BANNER performance 3

Supplement: Contest types 3

Option A. Prediction of secure expert-annotated data (EXS) 3

Option B. Prediction of mechanical turk-annotated data (MT) 3

# I. Objective

**Ultimate:** We want to build a computational system that can read biomedical abstracts and identify which words or phrases correspond to diseases.

**Specific:**  
#1. Improve the performance of currently developed algorithm (BANNER) on expert-annotated dataset  
#2. Implement the capability for an algorithm to increase the prediction performance on expert-annotated dataset by learning on the data, annotated by mechanical turk crowd.  
#3. Collect the information of methods and approaches, applicable to the area

# II. Data

**Potentially available:**

* 700 abstracts, annotated by experts (**EX**, golden dataset). Publicly available.  
  (*Question*: *Any chance we can remove them from public domain?*))
* 1000 abstracts, annotated by mechanical turk users (**MT**). Secure.
* Instruction for annotation used by experts and MT annotators.  
  (*Question*: *Can you confirm that the instructions for experts and MT annotators were absolutely the same?*)
* Standardized input-output and performance evaluation system for the BANNER algorithm
* BANNER code itself

**Might be potentially requested:**

* 200-300 abstracts, annotated by experts (**EXS**, Secure golden dataset) or
* 1000? abstracts, annotated by mechanical turk users (much cheaper)

**Current performance benchmark:**

BANNER algorithm, after training on about 500 expert-annotated abstracts can predict expert annotations for another ~200 abstracts with precision and recall around 0.8 (cross-validated over the dataset).

# III. Contest options and suggestions

Our options are quite limited by a current exposure of the expert-annotated dataset. Given that, there are three approaches we can take:

1. Remove the dataset from open-source resources, and accept a risk that someone will mine the data out and submit an over-fitted solution. The possible ways to hedge the risk is to forbid using any pre-set parameters in the algorithm, and introduce an extra evaluation step, where the solutions are going to be tested for unjustified fitting. This is doable, but it may result in unhealthy spirit of the contest.

2. We can request for additional (secure) EXS dataset and use it for scoring (Option A in supplement)

3. We can make a bold move, turn the problem inside out, and use EX and MT to predict MT (Option B). By doing this, we:  
*a. Achieve the objective #1, as we can set the corresponding benchmark with BANNER algorithm and outperform it.  
b. Achieve the objective #2, as the solution will assess the differences between EX and MT datasets  
c. Achieve the objective #3, as the methods to predict MT should be similar to those to predict ES  
d. We already have a secure scoring dataset  
e. It does not require additional expert work  
f. We can run a bug race after, or pay extra to the winner of the Marathon, to redesign the solution to work other way around.*

**Suggestion:** I would personally recommend the option #3, collection of 1000 additional annotated abstracts from mechanical turk (*for more accurate scoring*), and using a comprehensive analysis of the BANNER (questions from the next section) to adjust properly the framework of and expectations from the competition.

This is something to think about!

# IV. BANNER performance

Since we already have a working algorithm, we can learn a lot about the properties of the dataset and, therefore, expectations for the contest, by testing BANNER output. So, I have a few questions to ask:

1. How long does it take to run a single 500-training-200-scoring test for BANNER? What are the memory/hardware requirements?
2. Have you tested the BANNER learning curve (F score vs training set size)?
3. Have you tried to train BANNER on MT only and predict ES? Train BANNER on ES and predict MT? Train BANNER on MT and predict MT?
4. What is the distribution (Mean average? Deviation?) of the F-scores in 500 training/200 testing cross-validation test?

# Supplement: Contest types

## Option A. Prediction of secure expert-annotated data (EXS)

**Data:** 700 EX + 300 EXS + 1000 MT

**Data partitioning (default):** 500 EX + 1000 MT – training subset;  
200 EX – provisional scoring; and 300 EXS – final scoring.

**Training:** The training subset + annotation and evaluation rules are provided to contestants + BANNER code are provided to contestants

**Testing procedure:** Provisional scoring procedure is based on provisional scoring dataset and can be used by contestants to evaluate their solutions while the contest is going. Final scoring procedure uses unexposed final scoring dataset and is performed only once, after the contest is completed.  
For scoring evaluation, the algorithm is applied to each of the scoring abstracts independently.

**Scoring** can be based on the F-score for each of the abstracts processed. Score is averaged between the abstracts in scoring subset (and multiplied by 1000). The scoring should be compared to BANNER score, achieved on the same test.

## Option B. Prediction of mechanical turk-annotated data (MT)

**Data:** 700 EX + 1000 MT (+ more MT to improve scoring)

**Data partitioning (default):** 700 EX + 500 MT – training subset;  
200(+) MT – provisional scoring; and 300(+) MT – final scoring.

**Training:** The training subset + annotation and evaluation rules are provided to contestants + BANNER code are provided to contestants

**Testing procedure:** Provisional scoring procedure is based on provisional scoring dataset and can be used by contestants to evaluate their solutions while the contest is going. Final scoring procedure uses unexposed final scoring dataset and is performed only once, after the contest is completed.  
For scoring evaluation, the algorithm is applied to each of the scoring abstracts independently.

**Scoring** can be based on the F-score for each of the abstracts processed. Score is averaged between the abstracts in scoring subset (and multiplied by 1000). The scoring should be compared to BANNER score, achieved on the same test.