Natural Language Processing (NLP) Workbench Web Services
Webinar
April 26, 2017

Questions and Answers

1. Will the workbench be deployed as a service for everyone to use, or can it be deployed at the local level to protect personal health information?
   The workbench will provide two deployment options: unsecured services to process publicly available data, and locally run services to protect confidential data.

2. Will the workbench allow users to plug in their own customized NLP components?
   Yes.

3. Is open source licensing being considered for the workbench?
   Not at this time.

4. Are you planning to have only one temporal analyzer or will you allow the user to choose from multiple analyzers?
   We will include multiple analyzers from which users may select the preferred tool for their use case.

5. Will the workbench allow the user to plug in particular lexicons and ontologies?
   Yes.

6. Will there be any evaluation capabilities so that performance of alternative modules can be compared?
   No, unless they come with the components we adopt.

7. We have a platform that uses the Galaxy workflow engine that enables users to create pipeline NLP web services (the LAPPS Grid). Are you familiar with Galaxy (originally developed for genomics)?
   Information about the LAPPS Grid was found through the literature review and environmental scan. We plan to explore its applicability for use in the clinical domain and our specific use cases.

8. Who are the targeted users for this workbench?
   There will be different levels of functionality based on the user’s familiarity with NLP pipeline and model development.

9. Are registrars in the National Cancer Institute’s program the target users? What about general cancer registrars?
   The NLP Workbench Web Services apply to any cancer registry program (hospital, region, state, and national). It will be available to other health domains and researchers also.

10. Will the workbench have a plugin style of architecture or be easily extendable?
    Yes, the workbench will be easily expandable and detailed instructions will be provided.

11. Is there a plan to develop a specification and/or API for the interface between the health system data which will allow sites downloading the workbench to know what web services they must implement to get the application to work?
    Yes.

12. Is the workbench going to pull data? If so, then the specification should be defined clearly up front.
    No. Users will have to send their own data to the service(s).

13. So, will this fit into the regular cancer registrar workflow or is it more for model testing?
    The plan is for this to be integrated into eMaRC Plus to process cancer pathology and biomarker reports. It will also be made available to any other software that needs to process text-based health data.

14. Is it fair to say that the workbench will create NLP web services (including a default web service) rather than a workbench that registrars would interact with?
    For the cancer use case, the workbench will be used to develop the NLP model and provide web services for eMaRC Plus to process cancer pathology and biomarker data and return coded cancer data. Cancer registrars will interact with the web services through eMaRC Plus to validate the model.
15. Why did you opt for a rule-based system?
   Initially, we will support the safety surveillance use cases using rule-based algorithms previously built and evaluated by the FDA or other groups. We will also examine the development of other machine learning or hybrid approaches.

16. Will there be opportunities for universities to donate machine learning-based systems?
   Yes, instructions on how contributions can be shared will be provided.

17. What’s the right way for the machine learning community to contribute?
   The machine learning community is encouraged to provide contributions on business and functional requirements for the NLP Workbench Web Service development at any time by sending e-mail to NLPWorkbench@cdc.gov or by participating in our quarterly calls. At the end of this project, steps will be clearly documented on how users can contribute their own models and data for inclusion in the workbench.

18. Which tool do you use to de-identify clinical text?
   The workbench will provide a list of de-identification tools and a compatible format output from the pipeline that can go into the chosen de-identification tool. De-identification of the data will be the user’s responsibility.

19. How will you address interoperability between services and tools? The GATE, UIMA, Stanford, OpenLP, LingPipe, and others use different formats.
   The workbench will use a common standard between the services and tools. We are exploring different options.

20. How will pipelines of web services be created if the services use different protocols (SOAP, REST, JSON-RPC)?
   The users will develop their own solutions.

21. Where will the services run? Sharing clinical data is always a problem, so would you plan on a way to build an application in the workbench, and then allow users to download it to run locally?
   They will run on our site, and also have the option to run locally.