BLAH4 Projects Index

Setup for Publication

- F1000 Hackathons Channel
  - Open Peer Review
  - Designed for short papers with “initial”/”prototype” idea/results
  - Under 1000 words is $150. Maybe include one overview figure.
  - Once enough reviewers have approved, it is Pubmed-indexed
  - Encourages edits based on reviewers (and they’re free)
  - https://f1000research.com/collections/hackathons
  - FAQs: https://f1000research.com/faqs

Annotation Framework Interoperability

- To establish interoperability between annotation frameworks and systems
- Annotation Frameworks
  - LAPPS Grid
  - AlvisAE
  - PubAnnotation
- Documentation
- BioTermHub - PubDictionaries (Fabio, Nico, Jin-Dong)
  - Nico is working on connection OGER and PubAnnotation
    - To call OGER from PubAnnotation so that users can annotate their corpora in PubAnnotation using OGER
  - Nico and Jin-Dong will work to connect BioTermHub with PubDictionaries
    - To establish portability of dictionaries in the two repositories.

Data-set Development

- SemRep LD (Tiffany, Lars Juhl Jensen, Jin-Dong, Maxat, Bill, Olivier)
  - See GitHub Wiki page for documentation
  - Hackathon Goals: Transform the National Library of Medicine’s Semantic Representation predications into an open linked data resource by mapping concept and relation annotations to Open Biomedical Ontologies.
    - Discuss solutions for UMLS licensing issues
    - Develop a plan for mapping ontologies
    - Finalize knowledge representation

        - Update code and generate triples

- SCIRT: (Ivo, Barbara, Larry, Kevin) Training neural net syntactic and semantic annotators using the CRAFT gold-standard corpus
    - Get access to project on GitHub
    - Download and generate descriptive statistics
    - Write up methodology
    - Write the paper
- AGAC (Jingbo, from HZAU)
    - Short intro: AGAC is short for Active Gene Annotation Corpus, which is a manual annotated corpus aiming to look for pharmacologically informative drug information by annotating 9 trigger noun labels and 3 trigger verb labels. The link of the AGAC page: http://xiajingbo.weebly.com/agac.html
    - File:https://drive.google.com/open?id=13J2mW5lJtHdkk_LuWjOEERepegjLoqtS
- Onto2Vec - Embeddings for AberOWL ontology terms based on Pubmed abstracts
    - Identify ontology classes in text using PubAnnotation, Pubdictionary or Tagger (by prof. Jensen) - Done!
    - Generate vector embeddings for ontology classes based on co-occurrence of the ontology classes in pubmed abstracts - Done
    - Use the embeddings to generate new annotations - Done
    - Project Documentation
    - Presentation
- Disease - gene - microbe
    - Annotated abstracts on Crohn's disease or on on *Staphylococcus aureus* for example:
      A. Abstract text https://organisms.jensenlab.org/document/8065002
      B. Disease mentions (Disease Ontology terms):
      https://diseases.jensenlab.org/document/8065002/annotations
      C. Organism mentions (NCBI Taxonomy taxa)
      https://organisms.jensenlab.org/document/8065002/annotations
      (pre-compiled annotations Open Annotation format)
      (consider also the tagger api)

      Update: Abstract dataset
      Upload to PubAnnotation as a Project
    - Report on the:
        - number of abstracts (132161)
        - number of entity types used (three for a start, organisms, disease, human genes)
        - number of annotations of each type
    - Proceed to comentioning score calculation (SPARQL)

    - Potential: provide input to Literature-based Knowledge Discovery with Relations (Jake)
    - Potential: consider the SPARQL queries for the PubAnnotation search pages

Update: worked with the 48 abstract that mention both Crohn's disease and *Staphylococcus aureus*
After annotating the 132K abstract dataset for 16hrs it is only 87% there and this does not include the JSON uploading to PubAnnotation

- Regulatory Network visualization - GitHub repository Documentation
    - From different sources, generate network visualisations (with a relevant tool : Cytoscape~~, Neo4j?~~) and merge them to build a global knowledge (with an evaluation about the usefulness).
    - Discussion and state-of-art about relevant tools or ways to visualise network (binary link - node and edge) taking into account the source of data (*i.e.* Cytoscape can allows give a weight on edge)
    - Discussion about the ways to manage data conflicts
    - Formatting data to be usable in the visualisation tool , and create individual networks
    - Evaluate the adding value of annotation (manual and/or automatic versus database)

System Development

- Literature-based Knowledge Discovery with Relations (Jake, Alex)
    - The goal is to predict edges in a knowledge graph that will appear in future publications
    - Public GitHub repository
    - Presentation for grand wrap-up session
    - Summary slide
- Machine-assisted Variant Curation
    - Machine-assisted Triage for Variant-related documents in SwissProt
    - Suggest machine-assisted triage workflow
    - Text classification method (CNN)
    - Apply it to other triage process of databases
- Context-aware co-occurrence scoring using CoCoScore
    - Proposal presentation: https://docs.google.com/presentation/d/1xffP9p_bt_5iirWmO3EYmm4QhDMPXDkqyrB2dVB3ZHo/edit#slide=id.g267b784ad7_0_9

    - Try scoring scheme on disease–phenotype associations as available in PubCaseFinder (Toyofumi + Alex - Toyofumi is preparing a dataset)
    - Extend scoring scheme for disease–gene associations beyond sentences towards paragraphs (Alex - not touched during hackathon but work for the next couple of weeks)
- Automatic annotation system for glycans (Shujiro, Kiyoko, Jin-Dong)
    - Try to add possible annotations to each glycans stored in GlyTouCan which is a repository database for glycan structures.
    - Dictionaries for glycobiology terms
    - A function to extract glycan annotations via PubAnnotation SPARQL interface in GlyTouCan
- Status of annotation service REST API (Jin-Dong)
    - To collect working REST call examples to get annotations from annotation services. The collection may be used for a quick reference to use the services, and also for a comparison purpose.
    - Documentation
- Mining linked annotations and Linked data (Senay)
    - Get annotated text (obtained from Jensen Lab (http://download.jensenlab.org))
    - Normalize text annotations (replacing text annotations with their database accession numbers)
    - Generate text embeddings
    - Index embeddings
    - Search for the similar vectors in Bio2Vec (http://bio2vec.net/query2.html)
    - Proof of concept: demonstrate the utility of Bio2Vec
        - case study: gene-chemical association

      See *Documentation* for details.

- Integration / Usage / Improvement of Bio Term Hub (github, License: BSD 2-clause)
    - To ease integration into automatic workflows, a REST API could be added. eCurrently BTH is accessible only via web interface, or (if installed locally) used through a command-line interface.
    - Currently the output is in tsv format. We propose adding a json-format output
        - evaluate different possible JSON representations and implement the best one
    - Contact: Fabio Rinaldi & Nico Colic
- OGER proposed activities:
    - Integrate OGER (through its Rest API) with another (better) visualization interface, e.g. PubAnnotations / TextAE
    - OGER supports BioC XML as both input and output, recently a JSON version of the BioC format has been defined. We propose to add support for this new format. While a possible approach would be to use the converter provided by the NCBI, it is preferable to use a solution with less overhead with respect to speed and memory consumption. Some useful repositories:

        - BioC-JSON
        - PyBioC: Python library for handling BioC
    - Only a fraction of the API's options is currently exposed in the web interface. Only allows specifying input documents through an ID or by typing/pasting plain text into a text box. The output is always an embedded HTML fragment with the annotations highlighted in color, which cannot easily be downloaded. We propose to extend the availables choices to the full range of input and output formats.
    - Contact: Fabio Rinaldi & Nico Colic

Pig-Easy Descriptive Statistics (PEDS)

This is probably a tautology, but I notice two trends in descriptive statistics of distributional phenomena:

- Statistics that people demonstrate are broadly **the same** across corpora, e.g. empirical demonstrations that Zipf's Law describes word frequency/rank relationships across sufficiently large collections of text
- Statistics that people demonstrate are unexpectedly **different** across different populations (textual genres, document sections, suicidal/non-suicidal people), e.g. proportions of first-person singular pronouns versus other pronouns

It strikes me that people write papers about one kind of statistic or the other. For example, people look at proportions of first-person singular pronouns without giving any indication of whether or not their data looks like a "sufficiently large" collection of text; people show Zipfian distributions without looking at anything other than word frequency/rank relationships. This seems like a problem, in that if we're only looking for things that we *expect* to be the same or *expect* to differ, we're probably missing things. Another problem, partially related: lack of documentation about how "basic" statistics are calculated makes cross-study comparison difficult or impossible. (For example: people frequently look at the proportion of first-person singular pronouns when they're interested in the linguistic correlates of mental illness, but they rarely specify what the denominator is. Total first-person pronouns? All pronouns? All pronouns, excluding epenthetic (non-referential) ones? They just don't say.)

A solution that I would propose for the problem of missing things because we only look for similarities *or* only look for differences: let's build a simple set of analysis tools that we could apply to arbitrary corpora. By "simple" I mean some combination of Perl scripts and R markdown documents, and by "analysis" I mean that the same bunch of scripts should give us values for both kinds of statistics. If we then provisionally define "arbitrary corpora" as "the variety of corpora that people are working on at BLAH," then we are in the perfect environment to test the analyses across a variety of corpora.

What I would suggest as far as the concrete approach would work like this:

- Tiffany and Jingbo provide statistical consultation for the set of things that we'll calculate
- Kevin and Nancy hack the inevitable reformatting code for all corpora, OR Nancy and Jin-Dong hack the code to download all corpora from PubAnnotation--seems like a reasonable and novel PubAnnotation use case
- Nancy and Prabha write the literature review
- Kevin write the R code for the calculations
- Kevin write the remainder of the paper
- Ivo and Barbara get the SCIRT corpus text files to Kevin, and we use those for initial development and testing
- Barbara reads successive drafts of the paper for intelligibility

Some examples of relevant work:

Cohen et al. (under review): distribution of different forms of the word "because" in French-language email messages about coordination of care for people with amyotrophic lateral sclerosis care (French has different words that all mean "because" and that are used differently depending on how much certainty the speaker thinks there is regarding the posited causal relationship), as compared to reference corpora, suggests that there is an enormous amount of speculation about why pretty much anything happens relative to this population. This raises some questions about whether or not we will actually be able to understand issues of coordination of care for people with amyotrophic lateral sclerosis, or at least whether or not we will be able to understand those issues by looking at the data in those email messages.

Cardoso et al. (in preparation): ontologies can differ quite a bit with respect to how much ambiguity there is in the terms that are associated with their concepts. This could have implications for named entity recognition and normalization of concepts from those ontologies. Are more-ambiguous terms more difficult to recognize/normalize than less-ambiguous terms? Do gross indicators of ambiguity suggest things that could be removed from concept "dictionaries" prior to doing concept recognition/normalization? Answers to these questions have implications for system-building.