Team #256 - CellFinder Text Mining Pipeline

Version 1.1 - July 29, 2013

This document describes the CellFinder text mining pipeline in the scope of curation of gene expression data in cells and anatomical parts. The CellFinder database[1] is a repository of cell research which aims to integrate data derived from many sources, such as literature curation and microarrays experiments. In scientific publications, curatable gene expression events correspond to text passages which show associations between a gene/protein, a certain cell/anatomical part and an expression trigger, i.e., a word which indicates that the event is taking place. The sentence below illustrates one such example (PMID 18989465):

On the other hand, the podoplanin expression occurs in the differentiating odontoblasts and the expression is sustained in differentiated odontoblasts, indicating that odontoblasts have the strong ability to express podoplanin.

A text mining pipeline has been developed for a curation of gene expression data in cells and tissues [1]. It relies on state-of-art freely available tools for the many phases of curation workflow (cf. Figure 1): document triage, recognition of a variety of entity types, event extraction and manual validation of the results. Most of these components have been evaluated on two manually annotated sets of 10 full text documents on kidney and human embryonic stem cell research [2, 1]. Named-entity extraction accuracy ranges from 40% to 90% (depending on the entity type) and performance of gene expression event extraction is around 50%.

In the current version of the pipeline, some of these components are not yet integrated, thus, some of the tasks need to be carried out separately. Currently, manual intervention from biocurators is only necessary for querying new documents for curation and validation of the extracted gene expression events. These two tasks and the tools associated to them are described below.

1 Document triage - MedlineRanker

Document triage is performed by MedlineRanker[2] [3]. It allows searching documents by querying PubMed using specified keywords or based on specified MeSH terms (cf. options 1 or 2 of "The Query Topic (The Training Set) is Defined By" in Figure 2). Instead of returning the publications derived from the query above, MedlineRanker uses their abstracts to internally train a classifier. This

[1] `http://www.cellfinder.org/`

[2] `http://cbdm.mdc-berlin.de/~medlineranker/cms/medline-ranker`

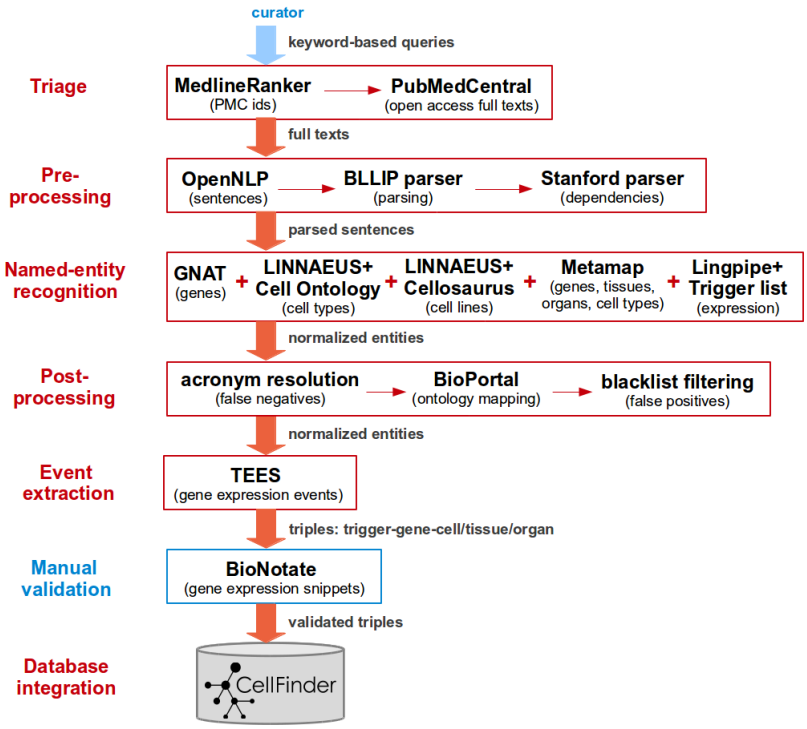

Figure 1: Text mining pipeline

derived classifier is then used to select and rank publications according to the options in "The Abstracts To Be Ranked (The Test Set) Are Defined By". Here we suggest choosing the "10,000 randomly chosen recent abstracts", i.e., relevant publications will be searched on a set of 10,000 randomly recent articles. Next, the user should click on the "Rank it!" button.

Next, a list of documents is shown to the user. The curator is free to have a look at the top-ranked results and perform new queries using different keywords in case he or she thinks that results are not relevant. We only request the list of top-ranked results to be sent to us, which can be easily downloaded from the link "Table of ranked PMIDs: tab-separated text table" at the bottom of the page (cf. Figure 3). The curator is free to perform more that one query and send the corresponding results to us. Finally, this step could also be performed internally by us, in case it is not possible to be carried out by the curator.

Provided the list of top-ranked results returned by MedlineRanker, we re-

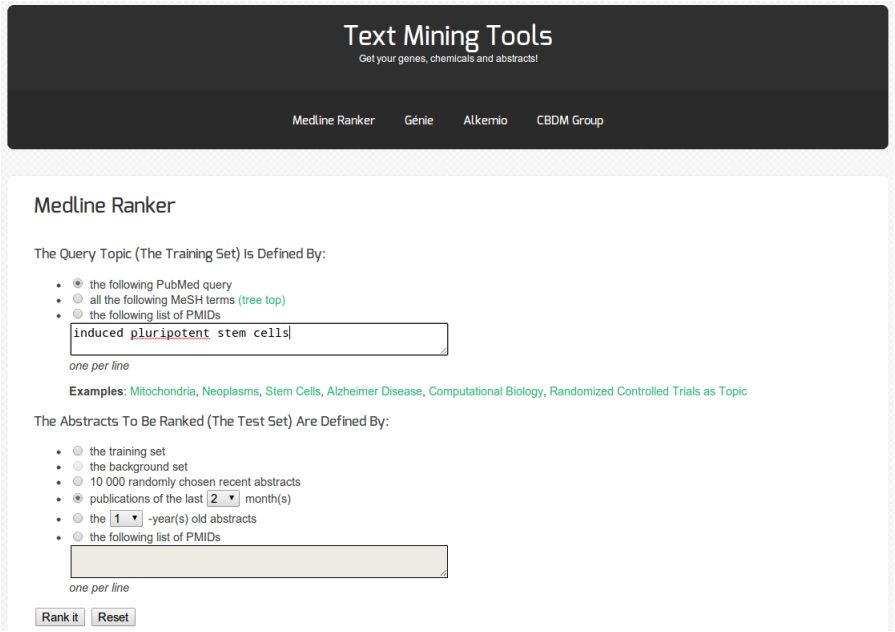

Figure 2: Screen-shot of the query page of MedlineRanker. Here we perform a search for publications related to "induced pluripotent stem cells."

trieve the corresponding abstracts and full texts (whenever available) from PubMed and PubMed Central, respectively. Next, these documents are processed through our text mining pipeline (cf. Figure 1). Then, the resulting gene expression events are loaded into Bionotate for manual validation and the corresponding URL will be sent to the curator. The processing of the documents can be carried out in one or two weeks, provided that the list of documents is not too long, i.e., up to 2000 documents. The workload is dependent on the number of distinct documents in the list, in case that the curator performs more than one query, and on the number of full papers available.

2 Data validation - Bionotation

Bionotate[3] [4] is a collaborative annotation tool developed for curation of relationships in biomedical texts. It is open-source code and configurable for other curation tasks. It works on a variety of browsers and has been tested in Firefox, Internet Explorer, Opera, Safari and Chrome.

We have configured Bionotate for curation of gene expression data as shown in Figure 4. Bionotate loads one gene expression event per time, which are randomly selected from the repository of extracted events. Pressing "F5" key makes the system lo load a new one, without making any changes to the currently one.

[3] `http://bionotate.sourceforge.net/`

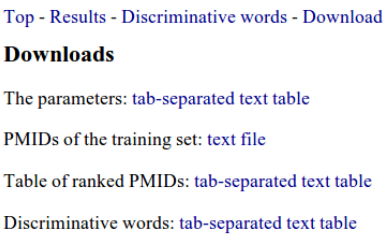

Figure 3: Screen-shot of the bottom of MedlineRanker results' page. The list of the 10,000 top ranked results should be sent to us.

When all snippets (gene expression) have been validated, the validation step is over and nothing more is loaded.

The following information is presented in Bionotate (cf. Figure 4):

- the PubMed identifier from where the data came from, along with a link to PubMed
- the entities of interest, i.e., an expression trigger, a gene/protein and a cell type, cell line or anatomical part
- a snippet of text, usually the sentence where the gene expression event was found and the two previous and following sentences, along with the entities of interest highlighted on the text (blue for gene/protein, green for the expression trigger and orange for the cell/anatomy)
- on the right of the text snippet, the highlighted entities are shown along with and "x", in case the curators needs to remove and add new one
- buttons for annotating new entities
- a question and a list of possible answers

The task is to check whether the entities have been correctly extracted, specially regarding the gene/protein and the anatomical part (cell/tissue/organ) and whether a gene expression is taking place. Curators are free to change the span of the entities. In this case, he or she should first remove the corresponding entity (by clicking on the "x" besides the text box) and then add the new one (by selecting a new text span and clicking on the corresponding entity button). The curators should only remove and add a new entity when the span of the automatically extracted one is partially correct. When the automatically extracted span is incorrect, and the curator notices that there is a correct entity somewhere else in the text (whether in the same sentence or not), he or she should not annotate this other entity, as it might have already appeared (or will appear later) in another text snippet. In summary, changes to the entities' span should only be carried out in the context of the "Entities of interest" listed above.

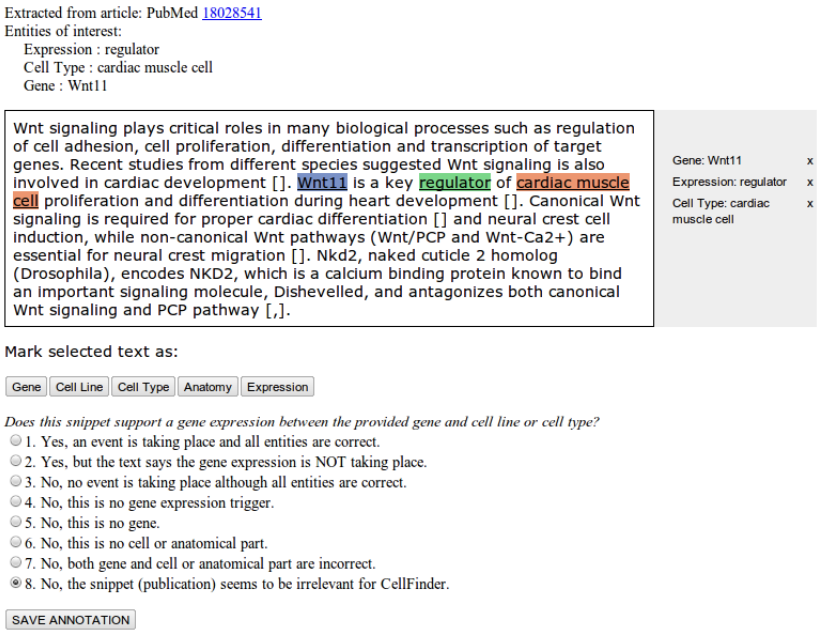

Figure 4: Screen-shot of Bionotate

Finally, one of the following answers should be chosen:

1. Yes, an event is taking place and all entities are correct.
2. Yes, but the text says the gene expression is NOT taking place.
3. No, no event is taking place although all entities are correct (default option).
4. No, this is not a gene expression trigger.
5. No, this is not a gene.
6. No, this is not a cell or anatomical part.
7. No, both gene and cell or anatomical part are incorrect.
8. No, the snippet (publication) does not seem to be relevant for CellFinder.

Answer 1 should be selected when all entities are correctly identified (whether automatically or after corrections) and they are indeed taking part in a gene expression event. This is a potential data to be integrated into CellFinder database as a positive expression level.

Answer 2 should also only be selected when all entities are correctly identified (whether automatically or after corrections) and when they are taking part in a gene expression event. The difference with **Answer 1** is that the text indicates that there is negation, i.e., in fact the gene/protein is not being expressed

in the specified anatomical part. Therefore, this is also a potential data to be integrated into CellFinder database, but this time as a negative expression level.

Answer 3 is only important as feedback for the event extraction component of the text mining pipeline. Here all entities are (fully or partially) correctly identified (whether automatically or after corrections) but there is no gene expression event taking place, i.e., both entities are being cited in another context. Therefore, this is not a potential data to be integrated.

Answers 4, 5, 6, 7 are important as feedback for the name-entity recognition component of the text mining pipeline and indicate whether the expression trigger, gene/protein, cell/anatomy or both of them, respectively, were incorrectly extracted (not even partially). Therefore, this is not a potential data to be integrated.

Finally, **Answer 8** is important as feedback for the document triage component of the text mining pipeline. This option should be selected when the text seems not be related to cell research and specifically to characterization (gene/protein expression) of cells and anatomical parts.

After choosing one of the options above and clicking on the "save annotation" button, an XML file is generated and saved in our server with the changes on the entities and the chosen answer. A new snippet is then loaded on screen and the validation process continues. Bionotate does not allow to download the generated XML with the answers. However, whenever requested by the curator, we are pleased to send them to him or her when the whole validation process is over.

Some examples of snippets for validation are available at: `http://141.20.31.85/cellfinder/`. The curators are free to annotate them as exercise. These snippets were derived from the experiments described in [1].

References

[1] Neves, M. *et al.* Preliminary evaluation of the cellfinder literature curation pipeline for gene expression in kidney cells and anatomical parts. *Database* **2013** (2013).

[2] Neves, M., Damaschun, A., Kurtz, A. & Leser, U. Annotating and evaluating text for stem cell research. In *Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012) at Language Resources and Evaluation (LREC) 2012*, 16–23 (Istanbul, Turkey, 2012).

[3] Fontaine, J.-F. *et al.* Medlineranker: flexible ranking of biomedical literature. *Nucleic Acids Research* **37**, W141–W146 (2009). `http://nar.oxfordjournals.org/content/37/suppl_2/W141.full.pdf+html`.

[4] Cano, C., Monaghan, T., Blanco, A., Wall, D. P. & Peshkin, L. Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *J Biomed Inform* **42**, 967–977 (2009).