# Semi-automatic curation of chronic obstructive pulmonary disease phenotypes using Argo

Riza Batista-Navarro, Jacob Carter and Sophia Ananiadou

National Centre for Text Mining, School of Computer Science
University of Manchester
{riza.batista,jacob.carter,sophia.ananiadou}@manchester.ac.uk

**Abstract.** Argo[1] is a generic text mining workbench that can cater to a variety of use cases, including the semi-automatic curation of information from literature. It enables its technical users to build their own customised solutions by providing a wide array of interoperable and configurable elementary components that can be seamlessly integrated into processing workflows. With Argo's graphical annotation interface, domain experts can then make use of the workflows' automatically generated output to curate information of interest. As part of our participation in the User Interactive Task of BioCreative V, we asked five domain experts to utilise Argo for the curation of phenotypes relevant to the chronic obstructive pulmonary disease (COPD). Specifically, they carried out three curation subtasks over passages drawn from full-text PubMed Central papers relevant to COPD. These include: (1) the markup of phenotypic mentions in text, e.g., medical conditions, signs or symptoms, drugs and proteins, (2) linking of mentions to relevant ontologies, i.e., normalisation, and (3) annotation of relations between COPD and other mentions. Analysis of the resulting annotations shows that an increase in throughput (9 vs. 14 curated passages per hour) was obtained with text mining-assisted curation. Inter-annotator agreement measured based on concept annotations was at an average F-score of 68.12%. To evaluate the performance of the automatic curation workflow, we compared the annotations it produced against those provided by one of the domain experts and obtained an F-score of 66.97%.

**Key words:** Biocuration, Concept annotation, Normalisation, Relation annotation, COPD phenotyping

## 1 Introduction

Chronic obstructive pulmonary disease (COPD) is a category of medical conditions characterised by blockage of the lung airways and breathing difficulties. In 2011, it was the third leading cause of death in the United States, and has been predicted to become the third one worldwide by 2030 [1].

Phenotypes are an organism's observable traits which help in uncovering the underlying mechanisms of a patient's medical condition. In the case of COPD,

[1] http://argo.nactem.ac.uk

disease and clinical manifestations are heterogeneous and widely vary from one patient to another. Methods for identifying phenotypes (i.e., COPD phenotyping) have thus been adopted to allow for the well-defined categorisation of COPD patients according to their prognostic and therapeutic characteristics.

The task of identifying phenotypes within narratives and documents, i.e., phenotype curation, is a widely adopted practice especially within the clinical community. As the amount of relevant textual data (e.g., clinical records and scientific literature) has continued to grow at an increasingly fast pace, substantial time and effort are required from human experts in curating phenotypic information. Aiming to alleviate this burden on human experts, we developed text mining workflows for semi-automatic phenotype curation in our Web-based workbench, Argo [6]. To demonstrate and evaluate Argo's suitability for the task, we participated in the User Interactive Task of BioCreative V (IAT), enlisting the help of five curators who carried out COPD phenotype curation. Results from the effort indicate that Argo shows promise as a phenotype curation tool.

## 2 System description and methods

In this section, we provide an overview of Argo's system features followed by a detailed description of how the curation tasks were conducted.

### 2.1 Features and functionalities

Argo is a generic text mining (TM) framework. Rather than catering to a specific application or use case, it enables its technical users to build their own customised TM solutions by providing a wide array of interoperable and configurable elementary components that can be seamlessly integrated into processing pipelines, i.e., workflows. We outline below the various features of Argo which enable its biocuration capabilities.

*Web-based availability.* Developed as a Web application, Argo does not require its users to perform any software installation, and can be accessed using any of the following browsers: Google Chrome, Mozilla Firefox and Safari. All workflows are executed on remote servers and can proceed even when users close the application. The interface displays a listing of a user's currently running executions to allow for progress monitoring.

*Library of interoperable components.* Key to Argo's processing capabilities is its continuously growing library of elementary processing tools. Owing to their compliance with the industry-accepted Unstructured Information Management Architecture (UIMA), these interoperable components can interface with each other and when combined into meaningful workflows, can form tailored TM solutions that address specific tasks. The components in the library are broadly categorised into three groups. *Readers* are for loading input data, e.g., document collections, either from a user's own files or from external resources (e.g.,

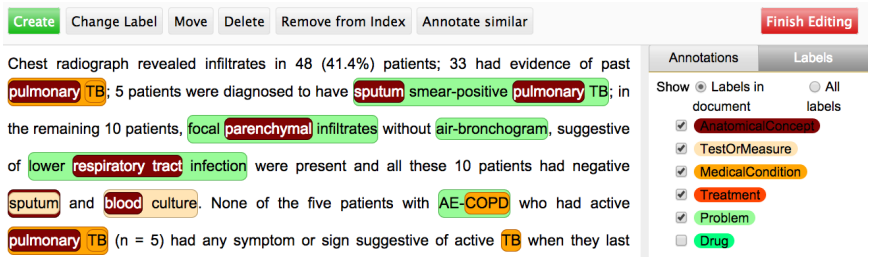



Fig. 1: Argo's annotation interface

PubMed). With readers for a variety of data formats such as plain text, tab-separated values (TSV), XML (e.g., BioC and XMI) and RDF, Argo enables its workflows to deserialise data from many publicly available corpora. Meanwhile, *Analytics* are implementations of various natural language processing (NLP) methods, and enrich input text with annotations at the lexical (e.g., lemmatisers), syntactic (e.g., tokenisers, dependency parsers) and semantic (e.g., named entity recognisers, concept normalisers and event extractors) levels. Finally, *Consumers* facilitate the serialisation of annotations to any of a user's preferred output formats (e.g., BioC, XMI, RDF and TSV).

*Workflow designer.* To support the creation of customised workflows out of the components previously described, Argo provides a block diagramming interface for graphically constructing TM workflows. A user designs a workflow by selecting components from the library, which will appear on the canvas as blocks. To define the processing sequence, the user arranges these blocks in the desired order and interconnects them using the available connection ports. Each of the components can then be customised with user-supplied parameter values. Guiding the user are detailed descriptions of each component's input and output types, as well as a panel that displays warning messages if problematic issues with a workflow have been detected.

*Manual and automatic modes of annotation.* One of Argo's available components is the Manual Annotation Editor which provides access to a graphical interface for manipulating annotations (Figure 1). To add new text span-based annotations, users highlight relevant tokens and assign suitable labels; annotated text spans are displayed according to an in-built colour-coding scheme. Structured annotations (e.g., relations, events) can be added by creating template-like structures and filling the slots either with primitive values or with any of existing text span annotations. Annotators can also remove annotations or modify the span, label or any other attribute value of existing annotations. Assignment of unique identifiers from external databases (e.g., for normalisation) is especially supported in Argo through an interactive utility for disambiguation that automatically retrieves a ranked list of matching candidates and displays further information coming directly from the relevant resource.

Argo supports different modes of annotation. For purely manual annotation, a workflow that consists only of a reader, the Manual Annotation Editor and any of the available consumers for saving annotations will suffice. In cases where text mining support is desired, we need to define to what extent we require the automation by incorporating chosen TM components into the workflow, before the Manual Annotation Editor. A curator can then use the Editor to revise the automatically generated annotations or supply his/her own new ones. It is also possible to visualise and revise annotated documents directly from a user's document space. This feature was incorporated into Argo to make it more convenient for annotators to review their previously annotated documents.

### 2.2 Task definition

As previously mentioned, the curation effort was comprised of three subtasks, described below.

*Markup of phenotypic mentions.* This subtask called for the demarcation of expressions denoting COPD phenotypes, which were also assigned semantic labels by the curators. Following the recent recommendation by Barker and Brightling [2] who argued that a multi-scale approach integrating information from various dimensions (e.g., gene, cell, tissue, organ) is necessary in order to fully understand a COPD patient's condition, we captured phenotypes falling under any of the following categories: medical condition, sign or symptom, protein and drug.

*Normalisation of mentions.* Many phenotypic concepts can be expressed in text in numerous ways. The phenotype pertaining to blockage of lung airways, for example, can take the form of any of the following variants and more: *airways are blocked*, *blocked airways*, *blockage of airways*, *airways obstruction*, *obstructed lung airways*. As a means for homogenising variants, the normalisation of surface forms to corresponding entries in ontologies was also required by our curation task. The following resources were leveraged: the Unified Medical Language System (UMLS) [3] for normalising mentions of medical conditions and signs or symptoms, UniProt [7] for proteins and ChEBI [5] for drugs.

*Relation annotation.* The last subtask involved the annotation of binary relations between a COPD mention and any other concept falling under our semantic categories of interest.

## 3 Results and Discussion

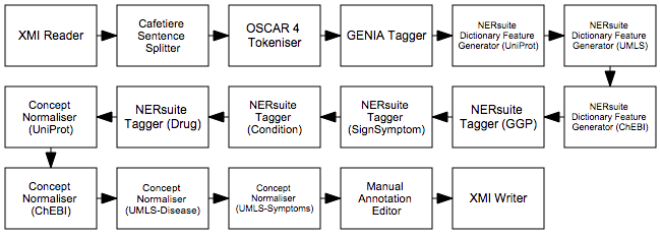
To assist our curators in accomplishing the tasks, task specifications and training material were provided. Firstly, annotation guidelines and detailed instructions were published as web pages, linked from Argo's main page[2]. A screencast

[2] http://argo.nactem.ac.uk/tutorials/curation-of-copd-phenotypes

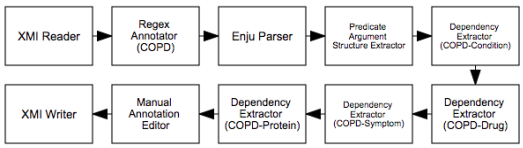
demonstrating the use of Argo's annotation interface was also prepared[3]. Furthermore, one-to-one tutorials were offered.

Based on the recommendation of the IAT organisers, the curators were requested to spend a total of at least four hours on this effort, distributed over two weeks. During the first week, they were asked to accomplish the first and second subtasks, i.e., marking up of phenotypic mentions and linking them to ontologies. The second week was then dedicated to the annotation of relations between concepts annotated during the preceding week. For each week, the curators provided their annotations in two modes. In the first mode, they were required to create annotations completely manually, i.e., without any TM support. In the second mode, meanwhile, they were given TM support in the form of automatically generated annotations. Depicted in Figures 2a and 2b are the TM workflows that we prepared in order to provide automatic curation support.

A corpus of 30 COPD-relevant PubMed Central Open Access papers that we have previously developed [4] was exploited in this effort. The corpus was split into two subsets with 15 papers each: one for training the text mining tools underpinning the automatic COPD phenotype curation workflows, and the other from which the documents for curation were drawn. Since the time constraints did not make the annotation of entire full-text papers feasible, we defined a document as a smaller chunk of text (e.g., section paragraphs according to each paper's metadata). Based on automatic random selection, 124 such documents were set aside for the curation task. The first 62 were used for purely manual curation while the remaining were exploited in the text mining-assisted mode of the task. All of the curators were asked to work on the same data set.




(a) Concept annotation workflow




(b) Relation annotation workflow

Fig. 2: Text mining workflows underpinning Argo's automatic COPD phenotype curation capabilities

[3] http://youtu.be/uOjwgmaXk00

For concept annotation (i.e., marking up phenotypic mentions and linking them to ontologies), the experts completed the curation of an average of nine passages in an hour. In the TM-assisted mode of concept annotation, the rate increased to an average of 14 passages per hour. Relation annotation was less time-consuming: in an hour, the curators annotated relations in 25 and 35 passages, in the non-TM and TM-assisted modes of annotation, respectively.

The curators were asked to annotate the passages in the same order that Argo displayed them, i.e., alphabetically by file name. In this way, even if the curators were carrying out their annotations at different rates (some curating more passages than the others within the allocated time), we were able to compile a corpus of 20 passages which were commonly annotated by all five curators. We estimated inter-annotator agreement (IAA) based on concept annotations (i.e., text span boundaries and semantic categories) manually produced for this set. We measured the F-score between each of the 10 pairs of annotators and obtained an average of 68.12% (lowest = 49.84%, highest = 82.78%).

Using the concept annotations (i.e., text span boundaries and semantic types) of the expert who voluntarily curated all of the 124 passages in the data set, we evaluated the performance of the Argo workflow which formed the basis of the text mining support provided to the curators. The overall precision, recall and F-score values obtained are 68.17%, 63.96% and 66.97%, respectively. These results are quite encouraging especially considering that the F-score (66.97%) is very close to the measured IAA (68.12%), indicating that our automatic concept annotation workflow performs comparably with human curators.

**Acknowledgments.** We are grateful for the annotation efforts of Laurel Cooper, George Gkoutos, Raymund Stefancsik, Loukia Tsaprouni and Nicole Vasilevsky.

## References

1. Chronic respiratory diseases. `http://www.who.int/respiratory/copd/en`, accessed: July 2015
2. Barker, B.L., Brightling, C.E.: Phenotyping the heterogeneity of chronic obstructive pulmonary disease. Clinical Science 124(6), 371–387 (2013)
3. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32(Database issue), D267–70 (2004)
4. Fu, X., Batista-Navarro, R., Rak, R., Ananiadou, S.: Supporting the annotation of chronic obstructive pulmonary disease (COPD) phenotypes with text mining workflows. Journal of Biomedical Semantics 6(1), 8 (2015)
5. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C.: The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Research 41(Database issue), D456–D463 (2012)
6. Rak, R., Rowley, A., Black, W., Ananiadou, S.: Argo: an integrative, interactive, text mining-based workbench supporting curation. Database : the Journal of Biological Databases and Curation 2012, bas010 (2012)
7. UniProt: Activities at the Universal Protein Resource (UniProt). Nucleic Acids Research 42(D1), D191–D198 (2014)