Development of bespoke machine learning and biocuration workflows in a BioC-supporting text mining workbench

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. As part of our participation in the Collaborative Biocurator Assistant Task of BioCreative V, we developed methods and tools for recognising and normalising mentions denoting genes/proteins and organisms. A combination of different approaches were used in addressing these tasks. The recognition of gene/protein and organism names was cast as a sequence labelling problem to which the conditional random fields algorithm was applied. In training our models, various lexical and orthographic features were extracted over the CHEMDNER GPRO and S800 corpora which were leveraged as gold standard data. Our feature set was further enriched with semantic attributes drawn from matches between mentions in text and entries in relevant dictionaries. In normalising recognised names, i.e., assigning resource identifiers to recognised names in text, the Jaro-Winkler and Levenshtein distance measures were used to estimate string similarity and -rank candidate matches. Integration of the various techniques and resources was facilitated by the Web-based Argo text mining workbench which allows for the straightforward construction of automatic text processing workflows. Upon using our training workflow to produce gene/protein and organism name recognition models and subsequently evaluating them, micro-averaged F-scores of 70% and 72.87% were obtained, respectively. Curation workflows applying our models on the provided BioC corpus of 120 full-text PubMed Central documents generated normalised named entity annotations which were serialised in the required BioC format.

Key words: Biocuration, Named entity recognition, Normalisation, Interoperability, BioC format

1 Introduction

The curation of biomedical information from literature is often a multi-step process consisting of several information extraction subtasks such as named entity recognition, normalisation and interaction extraction. An efficient approach to developing an end-to-end curation system therefore is to delegate different subtasks to several text mining experts. In this way, the burden of producing

an automatic curation system is distributed over different personnel, leading to more timely system completion. This, however, brings to the fore the need for interoperable annotations: the output produced by one contributor as part of one subtask needs to be easily ingestible by another party responsible for producing results in another subtask.

BioGRID is one example of a database whose content is largely reliant on information curated from the literature [5]. To support its curation, the Collaborative Biocurator Assistant Task of the Fifth BioCreative Challenge Evaluation has been dedicated to the coordination of different text mining teams as they simultaneously develop methods addressing certain subtasks out of the eight defined by the track organisers: (1) gene/protein named entity recognition (NER), (2) organism NER, (3) gene/protein name normalisation, (4) protein-protein interaction passage detection, (5) genetic interaction passage detection, (6) experimental method passage detection, (7) genetic interaction type passage detection, and (8) annotation visualisation. All developed tools were required to conform with the emerging XML-based BioC format [7], in order to facilitate the seamless exchange of annotations between different teams. This allowed participants of the track, henceforth referred to as the BioC track, to build their work based upon the results of other teams.

Contributing to this effort, we have chosen to participate in the first three subtasks focussing on the recognition and normalisation of gene/protein and organism names. In addressing the named entity recognition tasks, we took a supervised approach and developed models based on the conditional random fields (CRFs) algorithm [12]. For gene/protein and organism name normalisation, meanwhile, we estimated the similarity between textual mentions and name entries in external resources based on the Jaro-Winkler and Levenshtein distance measures [6]. The development of our methods, including the training of machine learning-based models, was carried out entirely using Argo, a Web-based workbench that facilitates the efficient creation of modular text mining workflows [15]. Owing to Argo's prevailing support for the BioC format, almost no effort was necessary in terms of serialising annotations according to the specifications set out by the track organisers.

2 Systems description and methods

In this section, we describe in detail the various techniques and resources that were exploited in order to address the NER and normalisation tasks. We illustrate in the succeeding section how we incorporated these methods into Argo workflows which ultimately produced the annotations contributed to the BioC track.

2.1 Pre-processing

All input documents were initially processed by a series of pre-processing tools. For detecting sentence boundaries, the LingPipe sentence splitter [3] was used.

Each resulting sentence was then segmented into tokens by the tokeniser packaged with the OSCAR4 tool [11]. The tokens were analysed by the GENIA Tagger [17] and assigned part-of-speech (POS) and chunk tags as well as their lemma.

2.2 Named entity recognition

The recognition of gene/protein and organism names was cast as a sequence labelling problem, i.e., the assignment of one of `begin`, `inside` and `outside` (BIO) labels to a sentence's tokens.

Our chosen implementation of CRFs is an in-house Java wrapper for the NERsuite package [4] which comes with modules for extracting dictionary features, training new models, and applying those models on documents. The following corpora served as our sources of gold standard annotations: CHEMDNER GPRO and S800 which contain gene/protein and organism name annotations, respectively. Although derived from patents rather than scientific literature, the CHEMDNER GPRO corpus [1] was selected as our gold standard corpus for gene/protein NER due to the large number of documents and annotations it contains. Each of its training and development subsets contains 7,000 documents, with the former having 6,877 annotations and the latter with 6,263. The S800 corpus, meanwhile, consists of 800 PubMed abstracts in which 3,708 organism names were manually annotated [14]. It is comprised of eight subsets with 100 abstracts each, pertaining to the following categories: bacteriology, botany, entomology, medicine, mycology, protistology, virology and zoology.

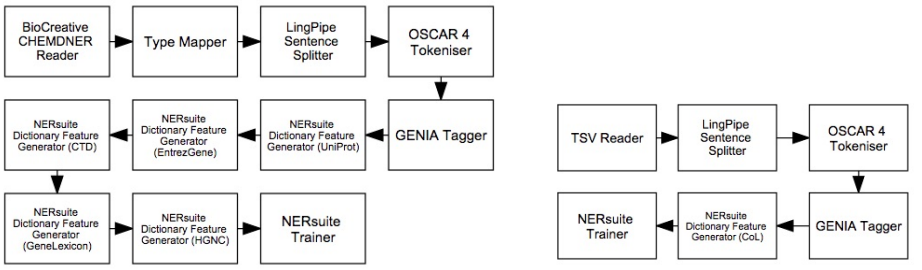
Each token was represented by a rich set of lexical, orthographic and semantic features, such as:

1. two, three and four-character n-grams
2. token, POS tag and lemma unigrams and bigrams within a window of 3
3. presence of digits or special characters
4. token containing only uppercase letters
5. word shape, with all of token's uppercase letters converted to 'A', lowercase letters to 'a', digits to '0' and special characters to '_'
6. matches against semantically relevant dictionaries

For the last type of features, we leveraged the following controlled vocabularies: UniProt [8], EntrezGene [13], the Comparative Toxicogenomics Database (CTD) [9], the HUGO Gene Nomenclature Committee (HGNC) database [10] and GeneLexicon [16] for gene/protein NER and the Catalogue of Life [2] for organism name recognition.

2.3 Normalisation

The normalisation subtask called for the automatic assignment of identifiers from the EntrezGene and NCBI Taxonomy databases to gene/protein and organism names, respectively.



(a) Gene/protein NER model training (b) Organism NER model training

Fig. 1: Training workflows

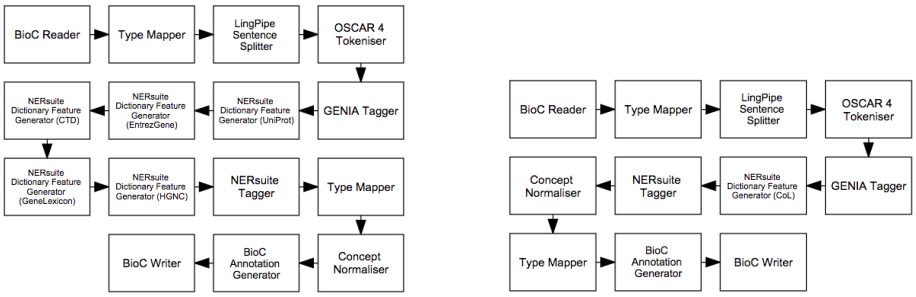
Although we planned to make use of semantics-based disambiguation methods (e.g., the use of organism information for gene/protein normalisation), in the interest of time, we took an unsupervised approach purely based on string similarity measures. Firstly, names (including synonyms) and corresponding identifiers from the databases of interest were collected. Each of the names was then converted to a canonical form based on the following series of steps:

1. converting all characters to lowercase
2. removal of stop words and punctuation
3. stemming of each remaining token
4. alphabetical re-ordering of tokens

The newly compiled dictionaries thus contain the equivalent canonical forms of names (and corresponding identifiers) from the databases. Each previously marked up gene/protein and organism name undergoes the same process of canonicalisation. The resulting canonical form is then used to query the relevant compiled dictionary for the most similar strings according to the Jaro-Winker distance measure. All entries in the dictionary whose similarity score is above a predefined threshold of 0.80 are considered candidates. Since multiple candidates having the same score were being returned, we additionally applied the Levenshtein distance measure in order to compute the similarity between the query name and a candidate. This allowed us to induce a more informative ranking of the candidates, from which the topmost result was considered as the best matching dictionary entry. The identifier attached to this entry is finally assigned to the name in question.

3 Results and Discussion

Argo facilitated the seamless integration of the above-described techniques and resources into text mining workflows. Figures 1a and 1b depict the workflows for training gene/protein and organism NER models, respectively. The first component of each of the workflows reads in documents containing gold standard annotations.



(a) Gene/protein NER and normalisation

(b) Organism NER and normalisation

Fig. 2: Curation workflows

In both cases, the NERsuite Trainer component makes use of the annotations provided by the preceding components (e.g., lemma, POS and chunk tags from the GENIA Tagger, dictionary matches from the NERsuite Dictionary Feature Generator) as features for training a CRF model.

To evaluate our methods, we trained NER models on training data subsets and subsequently evaluated them on held-out data. The model for gene/protein NER was trained on the CHEMDNER GPRO training corpus and then evaluated on the development set, as supplied by another BioCreative V track. In evaluating organism NER, meanwhile, we split the S800 corpus into two subsets of 400 randomly selected abstracts. One subset was used to train an initial CRF model while the other served as gold standard data for evaluation. Micro-averaged F-scores of 70% and 72.97% were obtained by the gene/protein and organism NER models, respectively. While the gene/protein NER model obtained balanced precision (67.71%) and recall (72.45%), the organism NER performs quite poorly in terms of precision (precision = 57.44%, recall = 100.00%). An immediate next step for us is to perform a detailed error analysis to resolve this issue.

In preparation for the application of our methods to the corpus of 100 full-text PubMed Central papers provided by the BioC track organisers, we retrained our organism NER model on the full S800 corpus. The trained models were then applied to the BioC-formatted corpus of 100 full-text PubMed Central articles (provided by the track organisers) using the workflows shown in Figures 2a and 2b. This was done by configuring the respective NERsuite Tagger components to specify paths to the pre-trained models. Based on the features extracted by the preceding components and presented to the model, the NERsuite Tagger generates named entity annotations. The Concept Normaliser, configured with a path to a pre-compiled dictionary of EntrezGene or NCBI Taxonomy entries, then assigns each of these named entities with the identifier of the best matching entry. The last component in each of the curation workflows serialises the generated annotations according to the BioC key file defined by the track organisers.

References

1. BioCreative V: Track 2 - CHEMDNER Patents. http://www.biocreative.org/tasks/biocreative-v/track-2-chemdner, accessed: July 2015
2. Catalogue of Life. http://www.catalogueoflife.org, accessed: July 2015
3. LingPipe 4.1.0. http://alias-i.com/lingpipe, accessed: July 2015
4. NERsuite: A Named Entity Recognition toolkit. http://nersuite.nlplab.org, accessed: July 2015
5. Chatr-aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., ODonnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., Tyers, M.: The BioGRID interaction database: 2013 update. Nucleic Acids Research (2012)
6. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. pp. 73–78 (2003)
7. Comeau, D.C., Islamaj Doan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T.C., Wu, C.H., Wilbur, W.J.: BioC: a minimalist approach to interoperability for biomedical text processing. Database 2013 (2013)
8. Consortium, T.U.: Activities at the Universal Protein Resource (UniProt). Nucleic Acids Research 42(D1), D191–D198 (2014)
9. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C., Mattingly, C.J.: The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Research (2014)
10. Gray, K.A., Daugherty, L.C., Gordon, S.M., Seal, R.L., Wright, M.W., Bruford, E.A.: Genenames.org: the HGNC resources in 2013. Nucleic Acids Research 41(D1), D545–D552 (2013)
11. Jessop, D., Adams, S., Willighagen, E., Hawizy, L., Murray-Rust, P.: OSCAR4: a flexible architecture for chemical text-mining. Journal of Cheminformatics 3(1), 41 (2011)
12. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
13. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 39(suppl 1), D52–D57 (2011)
14. Pafilis, E., Frankild, S.P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Arvanitidis, C., Jensen, L.J.: The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. PLoS ONE 8(6), e65390 (06 2013)
15. 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)
16. Tanabe, L., Wilbur, W.J.: Generation of a large gene/protein lexicon by morphological pattern analysis. Journal of Bioinformatics and Computational Biology 01(04), 611–626 (2004)
17. Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Bozanis, P., Houstis, E. (eds.) Advances in Informatics, Lecture Notes in Computer Science, vol. 3746, pp. 382–392. Springer Berlin Heidelberg (2005)