High-throughput, interoperability and benchmarking of text-mining with *BeCalm* biomedical metaserver

Miguel Madrid[1] and Alfonso Valencia[2]

[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, 28029 Madrid, Spain
[2]Life Sciences, Barcelona Supercomputing Centre, Barcelona, Spain.
mmadrid@cnio.es
alfonso.valencia@bsc.es

Abstract. Biomedical annotators are very specific tools applied to a highly complex field. Therefore, this kind of software suffers from an extreme complexity which impedes its usage. This complexity, which is reflected in usability problems, is the main cause of disuse, rejection and low impact. This document discusses several of these problems, as well as possible solutions. As a use case, the *NLProt* protein-names annotator and its benchmarking with the biomedical annotation metaserver *BeCalm* is analysed in detail.

Key words: *text-mining, biomedicine, BioCreative, metaserver, BeCalm, high-throughput, interoperability, benchmarking.*

1 Introduction

In the data-mining age, text-mining is a field of highest interest because the largest part of digital information is stored as unstructured text. The complexity is proportional to the amount information that contains: several idioms/dialect/jargon, language focused on different audiences (formal, informal, technical, simplified), expressiveness (feelings, assessments, context changes, annotations), figures, images, etc. A great effort and extensive knowledge are required to produce tools to analyse and extract relevant information and transform it into more compact, efficient and reusable formats.
In addition, the complexity of text and the importance of the topics addressed make it difficult to interpret them. For instance, in the Life Sciences (Biology, Molecular Biology and Biomedicine) the creation and usage of this kind of tools and systems increases constantly. However, there are no standards or shared evaluation criteria to establish quality measures of high-throughput, interoperability, scalability and reproducibility [1, Krallinger, M., & Valencia, A. (2005)]. This is why initiatives such as *BioCreative*, evaluation platforms such as *BeCalm*, and biomedical annotators such as *NLProt* are critical and of utmost importance.

1.1 BioCreative

BioCreative (Critical Assessment of Information Extraction systems in Biology) is an initiative in the text-mining field that was conceived to provide collaborative solutions to the problems arising in the construction and use of information extraction systems in the Life Science domain [2, Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005)] [3, van der Vet, P. E., van Ommen, G. J., Nijholt, A., & Valencia, A. (2001)].
Rather than a challenge, BioCreative is an effort to improve comparison methods, which define new resources for developing solid and effective gold standards by database curators and domain experts.

1.2 BeCalm

BeCalm (Biomedical Annotation Metaserver) was based on the *BioCreative* metaserver, a system for the remote administration of annotation servers [4, Leitner, F., Krallinger, M., Rodriguez-Penagos, C., Hakenberg, J., Plake, C., Kuo, C. J., ... & Johnson, C. A. (2008)] [5, Krallinger, M., Erhardt, R. A. A., & Valencia, A. (2005)]. *BeCalm* covers continuous assessment, interoperability between services and standard evaluation measures. Furthermore, it offers the possibility of including annotation web services as well as other resources to the final users.

1.2.1 TIPS

TIPS (Technical interoperability and performance of annotation servers) is a specific task of the *BioCreative* V.5 competition, which is evaluated by the *BeCalm* metaserver [6, Pérez-Pérez, M., Pérez-Rodríguez, G., Blanco-Míguez, A., Fernández-Riverola, F., Valencia, A., Krallinger, M., & Lourenco, A.]. This task is focused on the most technical part of text-mining (seconds per document, mean annotations per document, mean time in seconds to seek annotations and mean time in seconds per document volume) and it consists of providing a *REST API* for the annotator, which receives and answers requests from *BeCalm* in a number of given formats (*JSON*, *XML*, *TSV* or *BioC*).

1.3 NLProt

NLProt [7, Mika, S., & Rost, B. (2004)] is a tool for extracting and tagging protein-names in text by the combination of a dictionary added to a *Support Vector Machine (SVM)* [8, Tong, S., & Koller, D. (2001)] and linking them to their *UniProtKB/TrEMBL* identifications. Furthermore, *NLProt* is also able to tag tissues and species.

2 System description and methods

The requirements for the task consist of an annotation server accessed by a *REST API*. The aim is to measure three technical key aspects of annotation systems:

High-throughput: capacity to deal with large document volumes.
Interoperability: flexibility to handle different types of input and output.
Benchmarking: performance (mainly in speed) in different tasks.

2.1 BeCalm API

The *BeCalm* metaserver provides a very simple *REST API* for linking the annotator, comprising the following methods:

- *updateServerState*: The metaserver keeps continuous contact with the annotation server to analyse the behaviour and evolution of workload assigned to different subtasks.
- *getAnnotations*: The metaserver sends requests with a list of documents to annotate, including relevant meta information, such as document source and maximum execution time. Additionally, the method allows the incorporation of custom parameters.
- *saveAnnotations*: The annotation server returns annotations, complying strictly with the information provided by the *getAnnotatons* request.

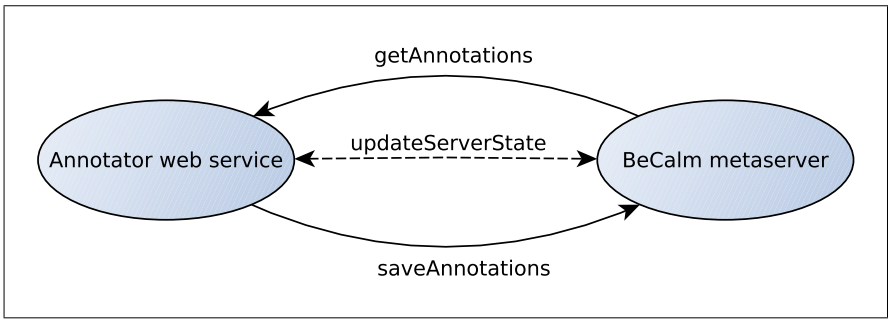

Fig. 1: BeCalm API diagram

2.2 NLProt

NLProt is an exceptional protein-name annotator. When considering partially tagged names as errors, *NLProt* still reached a precision of 75% at a recall of 76%. Nevertheless, despite its performance in tagging abstracts (see maximum times table [Subsection 3.1.2]), *NLProt* is not a command line tool designed for high-throughput batch processing and interoperability. In fact, *NLProt* does not process volumes of documents (all abstracts and titles have to be included in the same file) and both, the input and the output, have to be adapted to be used by other tools. Because of this, we developed the following wrapper solution:

2.2.1 Wrapper solution

Due to the above difficulties. A wrapper was developed with the programming language *Crystal*[1] with the help of the web framework *Kemal*[2], in order to facilitate the creation of the *REST API*. This wrapper solved all demands of the task [Fig. 2].

- *High-throughput*: The tool was prepared for massive batch processing. Moreover, an even-loop with fibres (as system threads but lighter and cooperatives) was implemented over the libevent library[3].
- *Interoperability*: Both, input and output of titles and abstracts, were adapted to the *JSON BeCalm* format.
- *Benchmarking*: With the above demands covered, the benchmarking of the annotator system is easily handled by the BeCalm metaserver.

[1] `https://crystal-lang.org/`
[2] `http://kemalcr.com/`
[3] `http://libevent.org/`

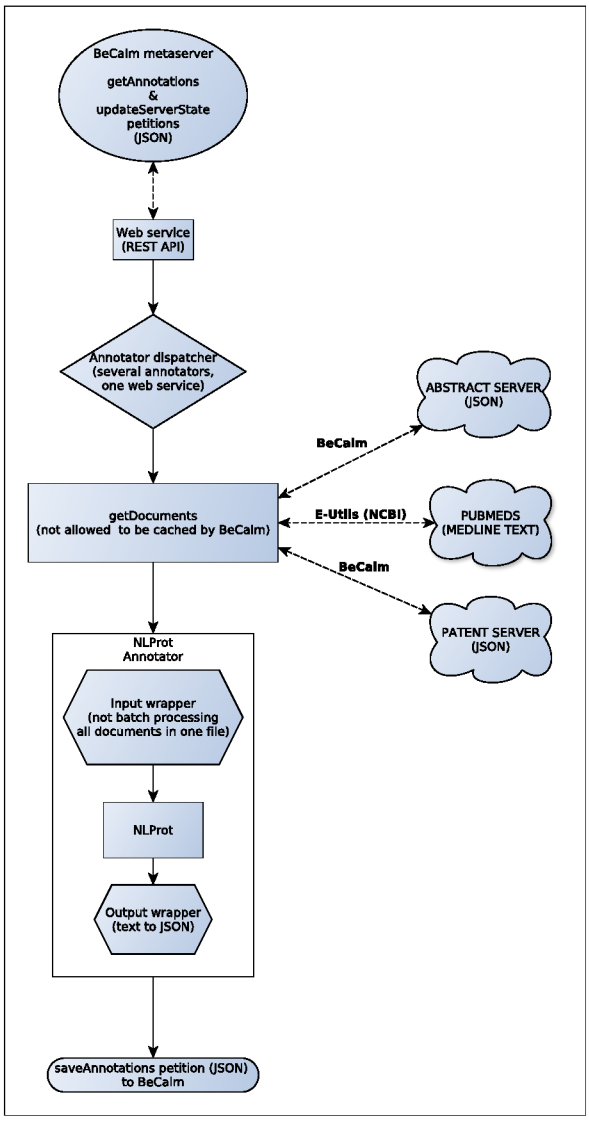

Fig. 2: Workflow of the web service annotator

3 Discussion

A virtual machine with 2 CPUs, 1GB of RAM and 5GB of hard drive was used for the task. Maximum times are due to the long stack of requests processed one by one considering hardware constraints.

3.1 Results

Results with all measures collected by the *BeCalm* metaserver.

3.1.1 Benchmarking by server

The *BeCalm* metaserver requires documents from three different servers preferably by *POST* requests (only way to download a large number of documents per volume). *BeCalm* monitorizes continuously maximum, minimum and average annotation times.

- *Patent*: server provided by *BeCalm* [Fig. 3].
- *Abstract*: server provided by *BeCalm* [Fig. 4].
- *Pubmed*: *NCBI* server [9, Canese, K., & Weis, S. (2013)] accessed through *E-utilites* [10, Sayers, E. (2009)] [Fig 5].

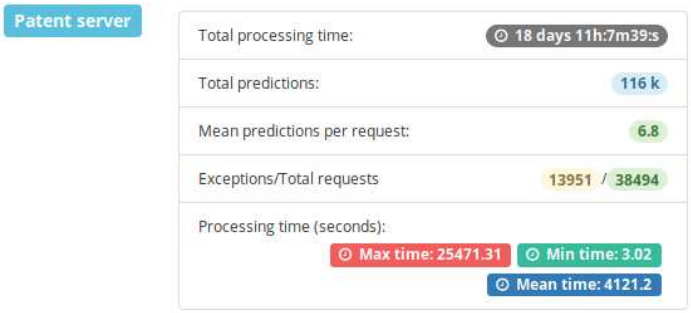

Fig. 3: *Becalm* Patent Server benchmarking

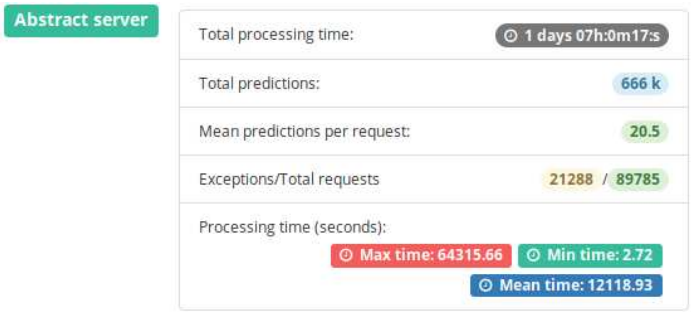

Fig. 4: *Becalm* Abstract Server benchmarking

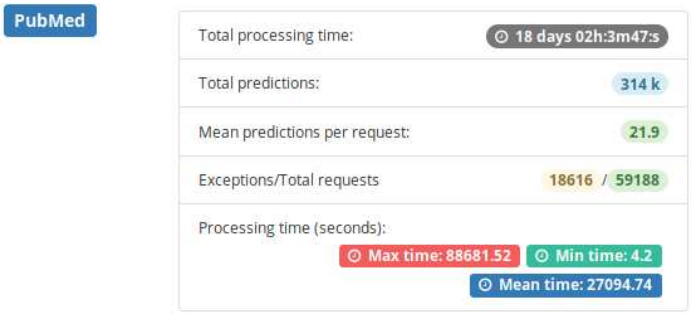

Fig. 5: *Becalm* PubMed Server benchmarking

3.1.2 Benchmarking by number of documents

BeCalm is focused on bursts of one document volumes. *NLProt* annotation time follows the trend of a linear function $f(x) = ax + b$ with $R^2 = 0.997$ [11, Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009)] with the number of documents per volume.

Documents (max time)	1	10	100	1000
getDocuments	0s	1s	2s	6s
NLProt	10s	22s	4m41s	45m3s
saveAnnotations	0s	1s	3s	18s

Table 1: Number of document per maximum time

3.1.3 General benchmarking

BeCalm also monitorizes a summary with average times from all document servers [Fig 6].

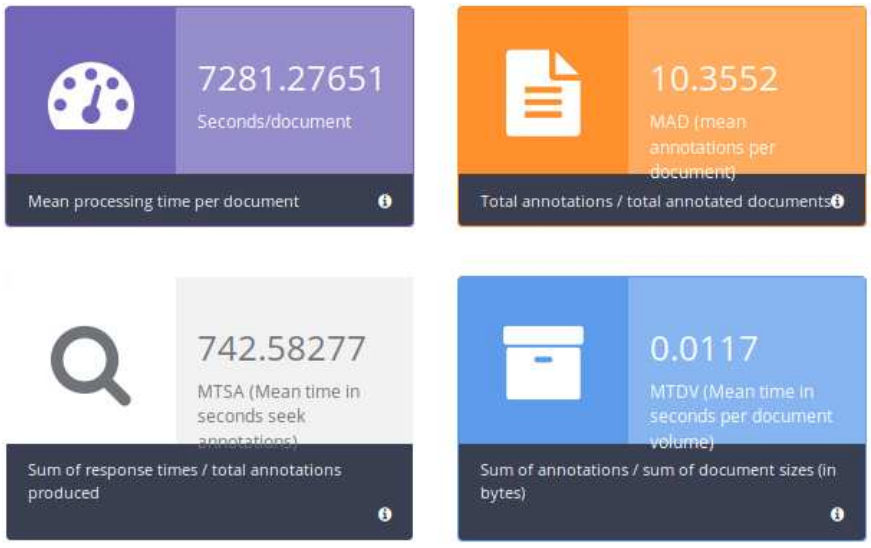

Fig. 6: *Becalm* general benchmarking

3.2 BeCalm API, web and feedback

The *BeCalm API* is simple but very powerful. It can be adapted easily with the custom parameters section of the requests. The web interface was designed to be minimalistic and user friendly, providing help and custom queries to test the annotation server and provide a summary of the query. In the future, it could be considered to add daily annotation statistics from other competitors.

4 Conclusion

The *BioCreative* task proves the critical importance of the evaluation of biomedical annotators with metaservers such as *BeCalm*, and demonstrates the need to adapt text-mining annotators to this type of standards. The evaluation should provide a qualitative assessment of the annotation servers in terms of functionality, as well as quantitative rating of technical key parameters. Evaluations have to be carried our against gold standards to obtain reliable and comparable measures for parameters such as as precision, recall or F-score.
To highlight, wrapper strategy over the annotator also has been tested successfully with the *Conditional Random Fields (CRF)* based [12, Okazaki, N. (2007)] *miRNA* entity tagger [13, Sammartino, J. C., Krallinger, M., & Valencia, A. (2016)] using the NERsuite toolkit [14, Cho, H. C., Okazaki, N., Miwa, M., & Tsujii, J. (2010)] for improving its behavior and to meet the demands (high-throughput, interoperability and benchmarking) of the task.

References

[1] Krallinger, M., & Valencia, A. (2005). *Text-mining and information-retrieval services for molecular biology.* Genome biology, 6(7), 224.

[2] Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics, 6(1), S1.

[3] van der Vet, P. E., van Ommen, G. J., Nijholt, A., & Valencia, A. (2001). *Information Extraction in Molecular Biology.*

[4] Leitner, F., Krallinger, M., Rodriguez-Penagos, C., Hakenberg, J., Plake, C., Kuo, C. J., ... & Johnson, C. A. (2008). *Introducing meta-services for biomedical information extraction.* Genome biology, 9(2), S6.

[5] Krallinger, M., Erhardt, R. A. A., & Valencia, A. (2005). *Text-mining approaches in molecular biology and biomedicine.* Drug discovery today, 10(6), 439-445.

[6] Pérez-Pérez, M., Pérez-Rodríguez, G., Blanco-Míguez, A., Fernández-Riverola, F., Valencia, A., Krallinger, M., & Lourenco, A. *Benchmarking biomedical text mining web servers at BioCreative V.5: the technical Interoperability and Performance of annotation Servers - TIPS track.* Proceedings of the BioCreative V.5 Challenge Evaluation Workshop., 12-21

[7] Mika, S., & Rost, B. (2004). *NLProt: extracting protein names and sequences from papers.* Nucleic acids research, 32(suppl 2), W634-W637.

[8] Tong, S., & Koller, D. (2001). *Support vector machine active learning with applications to text classification.* Journal of machine learning research, 2(Nov), 45-66.

[9] Canese, K., & Weis, S. (2013). *PubMed: the bibliographic database.*

[10] Sayers, E. (2009). *The E-utilities in-depth: parameters, syntax and more. Entrez Programming Utilities Help* [Internet].

[11] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). *Pearson correlation coefficient. In Noise reduction in speech processing* (pp. 1-4). Springer Berlin Heidelberg.

[12] Okazaki, N. (2007). *CRFsuite: a fast implementation of Conditional Random Fields.* 2015-03-24]. http://www. chokkan. org/software/crfsuite.

[13] Sammartino, J. C., Krallinger, M., & Valencia, A. (2016) *Annotation process, guidelines and text corpus of small non-coding RNA molecules: the MiNCor for microRNA annotations.*

[14] Cho, H. C., Okazaki, N., Miwa, M., & Tsujii, J. (2010). *NERsuite: a named entity recognition toolkit. Tsujii Laboratory, Department of Information Science*, University of Tokyo, Tokyo, Japan.