Benchmarking biomedical text mining web servers at BioCreative V.5: the technical Interoperability and Performance of annotation Servers - TIPS track

Martin Pérez-Pérez[1,2], Gael Pérez-Rodríguez[1,2], Aitor Blanco-Míguez[1,2], Florentino Fdez-Riverola[1,2], Alfonso Valencia[3,4,5,6], Martin Krallinger*[7], Anália Lourenço*[1,2,8]

[1]ESEI - Department of Computer Science, University of Vigo, Ourense, Spain
[2]CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310, Vigo, Spain
[3]Life Science Department, Barcelona Supercomputing Centre (BSC-CNS). C/Jordi Girona, 29-31, E-08034 Barcelona, Spain
[4]Joint BSC-IRB-CRG Program in Computational Biology. Parc Científic de Barcelona. C/ Baldiri Reixac 10, E-08028 Barcelona, Spain
[5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig de Lluís Companys 23, E-08010 Barcelona, Spain
[6]Spanish Bioinformatics Institute INB-ISCIII ES-ELIXIR, Madrid 28029, Spain
[7]Biological Text Mining Unit, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, C/Melchor Fernández Almagro 3, E-28029 Madrid, Spain

[8]CEB - Centre of Biological Engineering, University of Minho, Braga, Portugal

{martiperez, gaeperez,riverola,analia}@.uvigo.es; alfonso.valencia@bsc.es, mkrallinger@cnio.es

Abstract. The TIPS track consisted in a novel experimental task under the umbrella of the BioCreative text mining challenges with the aim to, for the first time ever, carry out a text mining challenge with particular focus on the continuous assessment of technical aspects of text annotation web servers, specifically of biomedical online named entity recognition systems.

A total of 13 teams registered annotation servers, implemented in various programming languages, supporting up to 12 different general annotation types. The continuous evaluation period took place from February to March 2017. The systematic and continuous evaluation of server responses accounted for testing periods of low activity and moderate to high activity. Moreover three document provider settings were covered, including also NCBI PubMed. For a total of 4,092,502 requests, the median response time for most servers was below 3.74 s with a median of 10 annotations/document. Most of the servers showed great reliability and stability, being able to process 100,000 requests in 5 days.

Keywords. TIPS; BeCalm metaserver; Document provider; Annotation server; Continuous evaluation; BioCreative; Text Mining

1 Introduction

There is an increasing demand in being able to effectively access, evaluate, compare, visualise and integrate multiple text mining systems in order to process natural language document collections. Several BioCreative tasks tried to promote the development of online text annotation servers [1–4]. In particular, the BioCreative Meta-Server was the first distributed prototype platform requesting, retrieving and unifying biomedical textual annotations [5]. Despite the relevance of those previous efforts, several crucial aspects have not been sufficiently or only partially addressed, including continuous evaluation, extraction of textual content from heterogeneous sources, harmonisation of multiple different biomedical text annotation types and visualisation and comparative assessment of automatic and manual annotations. This inspired the conception of the BeCalm Technical Interoperability and Performance of annotation Servers (TIPS) task for the BioCreative V.5 challenge.

This novel task focused on the technical aspects of making text-mining systems available, interoperable and continuously evaluating the underling named entity recognition web annotation servers. The participant annotation servers could be fully developed in-house or integrate/adapt third party recognition software as building block components. Furthermore, there were no restrictions in terms of named entity types/classes, thus covering entity type such as genes, proteins, chemicals, diseases or species among others.

In line with the efforts of ELIXIR/EXCELERATE in benchmarking the ELIXIR catalogue of methods and the OpenMinted interoperability specifications (http://openminted.eu/), both a minimal set of functional specifications (metadata info) and the use of a common communication protocol for serializing and distributing text annotations were reinforced. Specifically, the TIPS task considered three levels of evaluation: data level (i.e., data formats), technical level (i.e., stability and response time), and functional specification level (i.e., metadata requirements).

TIPS was supported by the *BeCalm* biomedical annotation metaserver (http://www.becalm.eu/) that enabled the continuous evaluation of annotation server performance as well as individual server monitoring by the participating teams. Annotation servers were asked to implement a Representational State Transfer (REST) API application that listens and responds to the requests made by the *BeCalm* metaserver. Annotation/prediction requests were issued on a regular basis, emulating different daily request loads during the months of February and March, 2017. Servers were forbidden to cache the documents, i.e. each document should be downloaded from the specified source whenever requested. Servers also should not cache the generated predictions, i.e., each document should be analysed for every request.

The aim of this paper is to describe the TIPS task and the specific support provided by BeCalm metaserver. The next sections present the architectural design of the metaserver, how the platform was utilised by the participants throughout the competition, and TIPS evaluation results.

2 BeCalm metaserver platform

The fundamental aim of the BeCalm biomedical annotation platform is to provide users with annotations on biomedical texts gathered from different systems. The platform is to be regarded as a distributed system requesting, retrieving and unifying textual annotations, to further deliver these data to the user at different levels of granularity.

For communication purposes, the system utilizes the REST API protocol [6]. The metaserver sends requests to annotate documents to all known/registered annotation servers. Once the annotation servers have finished processing the text, the predictions are returned to the metaserver and stored in its central repository. BeCalm REST API is publicly available at http://www.becalm.eu/api.

In assistance to TIPS competition (Figure 1), the BeCalm platform provided a user-friendly monitoring environment, where participating teams could manage annotation servers and examine their performance throughout the TIPS competition. Moreover, this monitoring environment offered participants the possibility of testing communication between the metaserver and the server, so that they could acquire insights on possible server improvements.

Regarding TIPS administration and functioning, the BeCalm platform enabled the registration of participants, the scheduling of annotation/prediction requests for continuous evaluation, the systematic calculation of server performance metrics, and a detailed log of events at both metaserver and server levels.

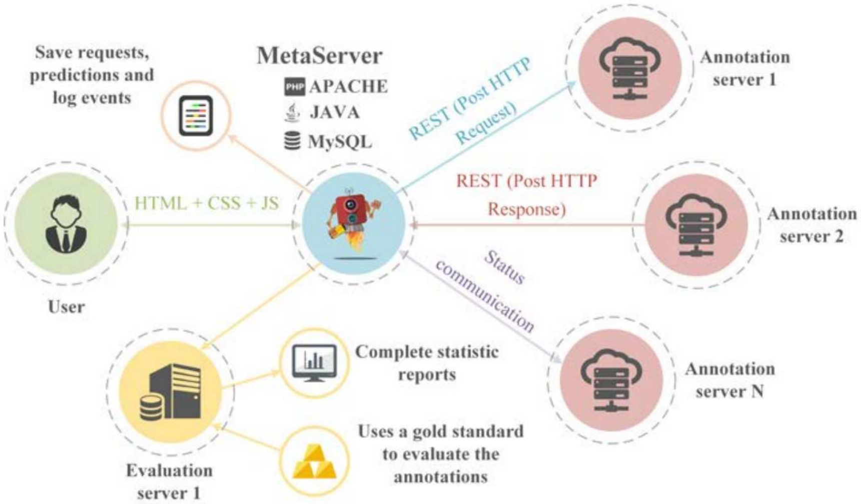


Figure 1: General description of BeCalm metaserver support to TIPS competition.

BeCalm interface is based on the open source CakePHP framework [7] and involves mainstream Web user-system interaction technologies, such as HTML5 (http://www.w3.org/TR/html5/), CSS3 (http://www.css3.info/), Ajax and JQuery technologies (http://jquery.com/).

3 TIPS competition

TIPS competition evaluates the technical aspects of making available and evaluating text annotation servers for continuous named entity recognition. At this first edition, servers were evaluated on the basis of single document requests.

3.1 TIPS evaluation levels

TIPS evaluation encompassed data format considerations, technical metrics and functional specifications. At the data level, evaluation addressed the ability of annotation servers to return NER annotation results as structured data, represented in one or several of the following formats XML/BioC, JSON/BioCJSON or TXT/TSV. The ability to retrieve and process documents from different providers (i.e., patents server, abstract server, and PubMed) was also examined.

Stability and response time were at the core of technical assessments. Stability metrics aimed to describe server ability to respond to continuous requests, to respond within a stipulated time window, and to provide updated server status information. Conversely, response time statistics described the time taken by the annotation server to respond to a request, measured in terms of the number and contents of the requested documents and the volume of predictions returned.

Functional specifications were inspired by the OpenMinTeD interoperability project (http://openminted.eu/). Server registration encompassed mandatory, recommended and optional metadata. Mandatory metadata included server name, institution/company, server administrator, programming language (main language, if using several), integration of third-party recognition software, recognised annotation types (e.g., chemical entities, genes, proteins, diseases, organisms, cellular lines and types, and mutations), supported annotation formats (e.g., XML/BioC, JSON/BioCJSON or TXT/TSV) and version control. Software license, specification of third-party recognition software (if any), dedicated vs shared server, and relevant publications were considered recommended metadata. Optionally, teams could also provide details on server operating system, distributed processing, and hardware characteristics (i.e., number of processors and RAM information).

3.2 TIPS evaluation metrics

Traditional annotation quality metrics (e.g., precision, recall, and F-score) were not part of TIPS evaluation. Rather, this novel task only evaluated performance metrics, namely reliability indicators and performance indicators (Table 1).

The mean time between failures (MTBF) and the mean time to repair (MTTR) are the key reliability indicators. Conversely, the mean annotations per document (MAD), the mean time per document volume (MTDV), the mean time seek annotations (MTSA), and the average response time (ART) are the key performance indicators.

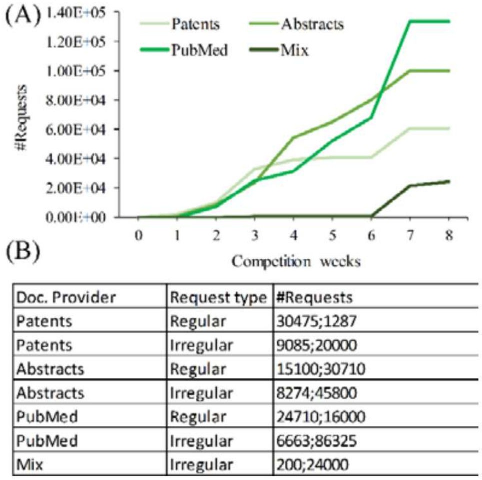
Table 1. Description of TIPS evaluation metrics.

Name	Equation	Description
MTBF	$(\sum(start\ of\ downtime(failure\ n+1) - start\ of\ uptime(failure\ n)))/(number\ of\ failures)$	Average elapsed time between failures of an annotation server.
MTTR	$(\sum(end\ of\ downtime(n) - start\ of\ downtime(n)))/(number\ of\ failures)$	Average time required to repair a failure in an annotation server, i.e. the necessary time to start the server again when a period of downtime occurs.
MAD	$(total\ number\ of\ annotations)/(total\ number\ of\ responses)$	Number of annotations divided by the total number of responses.
MTDV	$(\sum response\ time)/(\sum document\ size)$	Average time that the server takes to annotate a document (i.e. answer a request) based on the sum of the document sizes (in bytes) for all responses.
MTSA	$(\sum response\ time)/(total\ number\ of\ annotations)$	Sum of the response times divided by the total number of annotations produced.
ART	$(\sum response\ time)/(total\ number\ of\ responses)$	Average time to respond to a request.

4 Results

A total of 13 unique teams participated in TIPS. The annotation servers support a total of 12 unique annotation types. The chemical and disease types are the annotation types with greatest support (10 and 9 servers, respectively). The maximum number of types supported by a single server was 10 (server 120). Also, servers are implemented in various programming languages, namely Java (the most recurring), C#, C++, Node.JS, bash, Ruby, Python, Crystal.

The evaluation period started at February 5th 2017 and ended March, 30th 2017. The aim was to perform a systematic and continuous evaluation of server response under a varied request workload. So, the scheduling of annotation requests accounted for periods of low activity and moderate to high activity as well as for the three document providers, including a mix of them (Figure 2).



Doc. Provider	Request type	#Requests
Patents	Regular	30475;1287
Patents	Irregular	9085;20000
Abstracts	Regular	15100;30710
Abstracts	Irregular	8274;45800
PubMed	Regular	24710;16000
PubMed	Irregular	6663;86325
Mix	Irregular	200;24000

Figure 2. Requests issued per document provider throughout the evaluation period. (A) The plot depicts request per competition weeks from February 2017 to March 2017. (B) Information about the number of requests issued in February and March (semicolon separated) per document provider and request type.

Final performance results are shown in Table 2.

Table 2. TIPS evaluation data. Bold data represents the top values for each metric.

ID	#Requests	#Predictions	MTSA	MTDV	MAD	ART	MTBF	MTTR
103	**3.19E+05**	6.70E+05	7.58E-01	1.32E-03	2.13E+00	1.61E+00	**4.58E+06**	**0.00E+00**
106	3.12E+05	4.07E+06	8.59E-02	9.42E-04	1.34E+01	1.15E+00	**4.58E+06**	**0.00E+00**
107	2.95E+05	1.14E+06	2.85E+02	1.00E+00	4.27E+00	1.22E+03	4.62E+05	2.23E+05
108	1.23E+05	0.00E+00	-*	3.03E-02	0.00E+00	3.63E+01	**4.58E+06**	**0.00E+00**
111	3.11E+05	5.59E+05	3.55E+02	6.48E-01	2.27E+00	8.06E+02	5.19E+05	2.12E+04
114	**3.19E+05**	4.78E+06	1.21E-01	1.48E-03	1.51E+01	1.82E+00	**4.58E+06**	**0.00E+00**
116	2.29E+05	2.31E+06	3.83E+02	7.55E+00	2.35E+01	9.01E+03	8.11E+04	4.65E+05
117	**3.19E+05**	7.13E+06	1.29E-01	2.38E-03	2.25E+01	2.90E+00	**4.58E+06**	**0.00E+00**
120	2.91E+05	**2.74E+07**	**1.37E-02**	1.15E-03	**1.01E+02**	1.39E+00	**4.58E+06**	**0.00E+00**
121	**3.19E+05**	3.30E+06	1.18E-01	9.96E-04	1.04E+01	1.22E+00	**4.58E+06**	**0.00E+00**
122	3.16E+05	4.42E+06	7.23E-02	**8.58E-04**	1.48E+01	**1.07E+00**	**4.58E+06**	**0.00E+00**
124	4.98E+04	2.98E+04	1.55E+01	4.49E-02	3.29E+00	5.14E+01	1.17E+06	6.09E+04
126	4.98E+04	3.22E+04	1.50E+01	5.00E-02	3.69E+00	5.58E+01	5.86E+05	8.98E+04
127	**3.19E+05**	2.79E+06	4.20E-01	3.07E-03	8.90E+00	3.74E+00	**4.58E+06**	**0.00E+00**
128	1.87E+05	8.57E+05	5.44E+02	6.35E+00	1.38E+01	7.52E+03	1.73E+05	1.47E+05

*This server provided empty prediction files for all requests.

Servers 103, 114, 117, 121 and 127 have processed the biggest number of requests (3.19E+05). Server 120 has generated the largest number of predictions (2.74E+07), with an average of 101 predictions per document (MAD). In average, each prediction for server 120 has been generated in 0.013 s (MTSA). The minimum processing time value (ART) was 1.07 s, and the minimum processing time per document volume (MTDV) was 8.58E-04 bytes/s (server 122). During the whole TIPS competition, 9 servers have operated uninterrupted. Among the rest, the server 111 had the smallest recovering score (MTTR) with a value of 5.8 h.

5 Discussion

Overall, server performance metrics are quite encouraging, for example, for a total of 4,092,502 requests, the median response time for most servers was below 3.74s with a median of 10 annotations per document.

In terms of document provider, the median response time was 2.85s for the patent server, 3.01s for the abstract server and 3.48s for PubMed. PubMed slightly higher times are justified by the need of retrieving the abstracts at the time of the request, i.e. depending on PubMed service. Most of the servers showed great reliability and stability. Most of them were able to process 100,000 requests, for different providers, in five days. Considering that many participants have stated that their servers could perform batch processing, this figure is very promising, because the volume of processed documents could grow easily to one million documents.

Following this development path, the next TIPS evaluation phases will address multi-document requests, stress server tests and full-text annotation requests.

6 Acknowledgment

We acknowledge the OpenMinted (654021) and the ELIXIR-EXCELERATE (676559) H2020 projects, and the Encomienda MINETAD-CNIO as part of the Plan for the Advancement of Language Technology for funding. The Spanish National Bioinformatics Institute (INB) unit at the Spanish National Cancer Research Centre (CNIO) is a member of the INB, PRB2-ISCIII and is supported by grant PT13/0001/0030, of the PE I+D+i 2013-2016, funded by ISCIII and ERDF.

REFERENCES

1. Krallinger M, Vazquez M, Leitner F, et al (2011) The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinformatics 12 Suppl 8:S3. doi: 10.1186/1471-2105-12-S8-S3
2. Krallinger M, Morgan A, Smith L, et al (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol 9 Suppl 2:S1. doi: 10.1186/gb-2008-9-s2-s1
3. Wiegers TC, Davis AP, Mattingly CJ (2014) Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. Database 2014:bau050-bau050. doi: 10.1093/database/bau050
4. Wei C-H, Peng Y, Leaman R, et al (2016) Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford). doi: 10.1093/database/baw032
5. Leitner F, Krallinger M, Rodriguez-Penagos C, et al (2008) Introducing meta-services for biomedical information extraction. Genome Biol 9 Suppl 2:S6. doi: 10.1186/gb-2008-9-s2-s6
6. Massé M (2012) REST API design rulebook. O'Reilly
7. Iglesias M (2011) CakePHP 1.3 application development cookbook : over 60 great recipes for developing, maintaing, and deploying web applications. Packt Pub