Olelo's named-entity recognition web service in the BeCalm TIPS task

Hendrik Folkerts and Mariana Neves

Hasso Plattner Institute, August-Bebel-Str. 88 Potsdam, 14482 Germany
{Hendrik.Folkerts@student.hpi.uni-potsdam.de
Mariana.Neves@hpi.de}
https://hpi.de/plattner

Abstract. Named-entity recognition (NER) is an important preliminary tasks in many text mining systems. However, few web services are currently freely available for use. The BeCalm TIPS challenge aims to evaluate web services for biomedical NER in terms of reliability and performance. We participated with our dictionary-based NER which is part of our Olelo question answering system. Since the start of the evaluation on March 2017, our web services has received more than 290,000 requests and returned more than 30 millions annotations for ten annotation types and for three types of documents (patents, abstracts and PubMed).

Key words: Named-entity recognition, dictionary-based matching, MeSH, UMLS, in-memory database, web services.

1 Introduction

Named-entity recognition (NER) is an important preliminary task in many text mining workflows [1]. Although there are plenty of tools available for extracting a variety of entity types, such as for species [4] or disease names [3], text miners still need to install and integrate various tools for the various entity types. Web services offers the ability to obtain annotations for documents without the need to deal with installation and functionality of individual tools.

The BeCalm TIPS (technical interoperability and performance of annotation servers) challenge [1] [5] aims to evaluate the performance of web services for biomedical NER. Previously, many challenges have evaluated the performance of NER taggers for various entity types in terms of F-measure, precision and recall, such as [7]. These metrics focus mainly on the quality of the annotations returned by the systems, i.e., the rate of the correct annotations returned by the tools as well as whether the tools missed correct annotations.

As opposed to these previous challenges, BeCalm focus on the technical aspects of NER taggers and uses different evaluation metrics, such as reliability and performance indicators. Web services need to comply with some requirements, such as to provide a REST interface and output in standard format (e.g., BioC

[1] http://www.becalm.eu/pages/biocreative

XML). Finally, services are also required not to cache neither the documents nor the previously predicted annotations. The BeCalm TIPS challenge took place from beginning of February 2017 and ran until April 2017. The challenge is composed of four phases, as described below:

- Phase 1 - one document per request: five requests of single documents (patents or abstracts). The documents need to be retrieved from the corresponding server and annotations should be returned.
- Phase 2 - stress test: simultaneous requests to check the whether systems are able to manage these without collapsing.
- Phase 3 - bulk processing: request of multiple and large documents to check robustness and scalability of the systems.
- Phase 4 - PubMed/PMC: requests for documents from patent servers, PubMed and PubMed Central (PMC) to check the flexibility of the system to retrieve and process documents from various sources.

In this work, we present our Olelo web service for biomedical NER. Olelo[2] is a question answering (QA) system for biomedicine [2] which includes many natural language processing (NLP) components, such as question understanding, document retrieval, NER, answer extraction and summarization. The system is built on top of an in-memory database (IMDB) and uses a dictionary-based NER approach, as previously described in our last participation in the BioASQ challenge [6].

This paper is structured as follow: next section describes details of our NER system, followed by the results that we obtained in the BeCalm TIPS challenge.

2 System description

In this section we describe the details of our web service that participated in the BeCalm TIPS challenge. The system relies on the NER components behind our Olelo QA system [6]. Figure 1 illustrates the many components in the service. Details for each of them is provided below.

The components of our workflow which are not included in the IMDB, i.e, JSON reader, document retrieval and BioC writer, are implemented in the Java programming language. We use the Spring framework to combine the different components of our application. The core of our system is an IMDB (SAP HANA) that runs in a machine with 120 cores and 2067Gb of main memory.

JSON reader. Requests from BeCalm TIPS are in JSON format and includes the document identifiers and databases.

[2] `http://hpi.de/plattner/olelo`

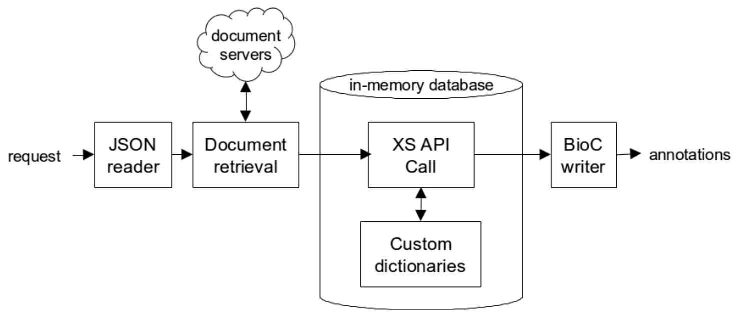

Fig. 1. Workflow of the web service.

Document retrieval. Given the document identifiers and documents databases extracted in the previous step, we proceed to retrieve the documents from the respective document servers. Participating systems need to provide support for four document servers: a patent server, an abstract server (both maintained by BeCalm), as well as PubMed and PubMed Central e-utils web servers. For each of these documents, we extract the title and the text of the abstract.

	PMID	TA_COUNTER	TA_TOKEN	TA_TYPE	TA_NORMALIZED	TA_SENTENCE	TA_OFFSET
29	24353217	29	drug	T121	C0013227	1	105
30	24353217	30	offenders	M01.142	offenders	1	116
31	24353217	31	offenders	T098	C0699726	1	116
32	24353217	32	This	T204	C1080058	2	127
33	24353217	33	specific	T080	C0205369	2	132
34	24353217	34	specific	T170	C1552740	2	132
35	24353217	35	review	V02.912	review	2	141
36	24353217	36	review	V02.600.500	review	2	141
37	24353217	37	review	T170	C0282443	2	141
38	24353217	38	review	T078	C1552617	2	141
39	24353217	39	pharmacological	T169	C0205464	2	158
40	24353217	40	interventions	T058	C0886296	2	174
41	24353217	41	at	T023	C0228317	2	194
42	24353217	42	reducing	T080	C0392756	2	197
43	24353217	43	drug use	T033	C2239127	2	206

Fig. 2. Screen-shot of some of the annotations for the sentence "This specific review considers pharmacological interventions aimed at reducing drug use and/or criminal activity for illicit drug-using offenders." (PMID 24353217).

XS API call. In this step we send the document content to the XS application of our IMDB. The XS application uses an built-in procedure from the IMDB which includes some preliminary NLP procedures such as language detection, sentence splitting, tokenization, stemming and part-of-speech tagging. The response of this API call includes all recognized named-entities. All documents sent to this

API are later deleted to comply with the requirement of the BeCalm challenge of not to keeping local copies of the documents. A screen-shot of some recognized named-entities is shown in Figure 2.

UMLS types	**BeCalm types**
Anatomic components	T017 (Anatomical Structure), T029 (Body Location or Region), T022 (Body System), T021 (Fully Formed Anatomical Structure), T018 (Embryonic Structure)
Cell line and cell type	T025 (Cell)
Chemical	T103 (Chemical), T120 (Chemical Viewed Functionally), T104 (Chemical Viewed Structurally)
Disease	T020 (Acquired Abnormality), T190 (Anatomical Abnormality), T049 (Cell or Molecular Dysfunction), T019 (Congenital Abnormality), T047 (Disease or Syndrome), T050 (Experimental Model of Disease), T033 (Finding), T037 (Injury or Poisoning), T048 (Mental or Behavioral Dysfunction), T191 (Neoplastic Process), T046 (Pathologic Function), T184 (Sign or Symptom)
Gene	T028 (Gene or Genome)
miRNA	T086 (Nucleotide Sequence)
Organism	T100 (Age Group), T011 (Amphibian), T008 (Animal), T194 (Archaeon), T007 (Bacterium), T012 (Bird), T204 (Eukaryote), T099 (Family Group), T013 (Fish), T004 (Fungus), T096 (Group), T016 (Human), T015 (Mammal), T001 (Organism), T101 (Patient or Disabled Group), T002 (Plant), T098 (Population Group), T097 (Professional or Occupational Group), T014 (Reptile), T010 (Vertebrate), T005 (Virus)
Protein	T116 (Amino Acid, Peptide, or Protein)
Subcellular structure	T026 (Cell Component)
Tissue and organ	T024 (Tissue), T023 (Body Part, Organ, or Organ Component)

Table 1. Mapping of UMLS semantic types to BeCalm annotation types.

Custom dictionaries. We compiled a custom dictionary based on MeSH terms and concepts from the UMLS database. Previously, we identified ten high-level nodes of the MesH tree and mapped these to UMLS semantic groups or types: A (Anatomy), B (Diseases), C (Species), D (Drugs), E01 (Clinical Diagnostics), G05.360.340.024.340 (Genes), D12.776 (Proteins), Z (Geographicals), E02 (Treatments), C23 (Symptoms). In order to account for the annotation types supported by BeCalm, we manually mapped these to the UMLS semantic types, as shown in Table 1. Currently, we are providing annotations for ten of the twelve annotations types supported by BeCalm TIPS. We do not support entity name normalization.

BioC writer. In this step we generate the BioC file with the corresponding annotations to the requested documents. We keep no copy of the extracted annotation from previous request, in order to comply with this requirement from the BeCalm TIPS challenge.

3 Evaluation

In this section we describe the metrics utilized by BeCalm TIPS for evaluation of the systems and we present the results that we obtained so far with our web service. These are the official results that are available to each participants in the BeCalm TIPS participant's area.

3.1 Evaluation metrics

As discussed earlier, BeCalm TIPS utilizes evaluation metrics which focus on the reliability and time performance of the systems instead of the quality of their annotations. More details are available in the BeCalm documentation [3].

Regarding reliability, BeCalm proposes two metrics:

- Mean time between failures (MTBF): average of the time elapsed between failures of an annotation server, i.e., the time elapsed between two consecutive failures.
- Mean time to repair (MTTR): time required to repair a failure, i.e., the time elapsed between a failure and the running of the system again.

Regarding performance metrics, the TIPS challenge proposes four metrics:

- Mean annotations per document (MAD): total number of annotations divided by the total number of documents.
- Mean time per document volume (MTDV): average time taken to annotate a document based on the sum of document sizes in bytes.
- Mean time seek annotations (MTSA): sum of response time divided by the total number of annotations returned.
- Average response time (ART): sum of response time divided by the total number of responses, i.e., average time per request/response.

3.2 Results

The values for some of the above metrics are available for participants in the participant's area, as well as additional statistics about the performance of the system. In this section we present the results we have obtained so far. Unfortunately, we only have access to our own results and cannot yet compare these with results from other participating systems.

Since the start of the evaluation period, Olelo has received a total of 291,452 requests (as of April 3rd, 2017). Table 2 summarizes the statistics of our results for the different document types.

[3] http://www.becalm.eu/files/material/BeCalm_TIPS.pdf

Docs.	Pred.	Req.	Mean pred/req	Excep.	Time (mean/min/max/)
Patent	6822k	90,213	77.9	2,339	1.23s/0.35s/102.84s
Abstract	14,833k	115,938	128.7	394	1.17s/0.36s/202.88s
PubMed	8,421k	85,301	121.5	15,049	1.99s/0.9s/37.08s

Table 2. For each document server, results are shown for the total number of predictions, the total number of requests, the mean number of predictions per request, the total number of exceptions and the processing time (mean, minimum and maximum). Values are as of April 3rd, 2017.

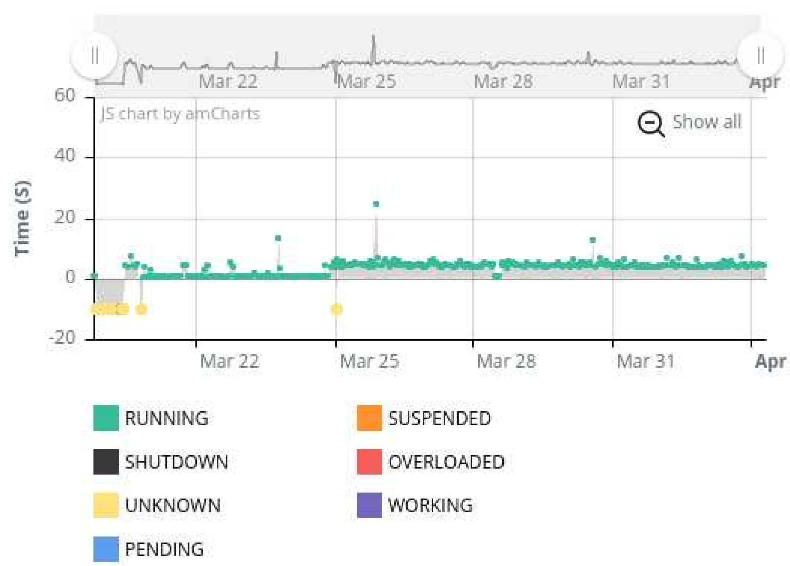

Fig. 3. Screen-shot of the historical time response of Olelo as shown in the BeCalm participant's area.

Regarding the metrics described above (cf. section 3.1), we obtained a MAD (mean annotations per document) of 100.8349, a MTSA (mean time seek annotations) of 0.01959 and a MTDV (mean time per document volume) of 0.0343. Additionally, the mean processing time per document was 1.39571 seconds.

Our historical time response is shown in Figure 3. So far, we obtained a total of 30,615 exceptions (as of April 3rd, 2017), i.e., cases in which our system did not successfully respond to the requests. We manually checked the last 50 exceptions returned by Olelo and the results are the following errors: 25 of type UNRECHEABLE_SERVER_Proxy, 23 of REQUEST_TRIES_OVERLIMIT, one NOT_VALID_BY_SCHEMA error and one REQUEST_CLOSED error. The most common error relates to DNS lookup failure for our server, an issue that we are currently looking into. The second most frequent exception indicates that our server reached the maximum number of attempts for accepting a request, which is probably related to the previous exception. Finally, the NOT_VALID_BY_SCHEMA error was due to errors in the XML DTD schema and the REQUEST_CLOSED exception seems to be due our system submitting more than one response for the same request.

4 Conclusions

We presented our participation on the BeCalm TIPS challenge for the evaluation of reliability and performance of web services for biomedical NER. We participated with the dictionary-based NER procedure behind our Olelo QA system, which relies on custom dictionaries from UMLS and MeSH. Our web service received so far more than 290,000 requests and returned annotations for almost 90% of these, totalizing more than 30 millions annotations.

Acknowledgments. We would like to thank Matthias Herzog for technical support and Milena Kraus for her support of mapping the semantic types.

References

1. Hirschman, L., Burns, G.A.P.C., Krallinger, M., Arighi, C., Cohen, K.B., Valencia, A., Wu, C.H., Chatr-Aryamontri, A., Dowell, K.G., Huala, E., Loureno, A., Nash, R., Veuthey, A.L., Wiegers, T., Winter, A.G.: Text mining for the biocuration workflow. Database 2012 (2012)
2. Kraus, M., Niedermeier, J., Jankrift, M., Tietböhl, S., Stachewicz, T., Folkerts, H., Uflacker, M., Neves, M.: Olelo: a web application for intuitive exploration of biomedical literature. Nucleic Acid Research Web Service issue (accepted) (2017)
3. Leaman, R., Islamaj Doan, R., Lu, Z.: Dnorm: disease name normalization with pairwise learning to rank. Bioinformatics 29(22), 2909–2917 (2013)
4. 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)
5. Pérez-Pérez, M., Pérez-Rodríguez, G., Blanco-Míguez, A., Fdez-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. In: Proceedings of the BioCreative V.5 Challenge Evaluation Workshop. pp. 12–21 (2017)
6. Schulze, F., Schler, R., Draeger, T., Dummer, D., Ernst, A., Flemming, P., Perscheid, C., Neves, M.: Hpi question answering system in bioasq 2016. In: Proceedings of the Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics. pp. 38–44 (0 2016)
7. Smith, L., Tanabe, L., Ando, R., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W., Hunter, L., Carpenter, B., Tsai, R., Dai, H.J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Mana-Lopez, M., Mata, J., Wilbur,

W.J.: Overview of biocreative ii gene mention recognition. Genome Biology 9(Suppl 2), S2 (2008)