

Assessment of NER solutions against the first and second CALBC Silver Standard Corpus.

Submitted by [rene](#) [1] on Mon, 2012-01-23 09:41

- [Bioinformatics](#) [2]
- [Text Mining](#) [3]

Title	Assessment of NER solutions against the first and second CALBC Silver Standard Corpus.
Publication Type	Journal Article
Year of Publication	2011
Authors	Rebholz-Schuhmann, D. [4], A. Jimeno Yepes [5], C. Li [6], S. Kafkas [7], I. Lewin [8], N. Kang [9], P. Corbett [10], D. Milward [11], E. Buyko [12], E. Beisswanger [13], K. Hornbostel [14], A. Kouznetsov [15], R. Witte [16], J. B. Laurila [17], C. J. O. Baker [18], C. - J. Kuo [19], S. Clematide [20], E. Rinaldi [21], R. Farkas [22], G. Móra [23], K. Hara [24], L. I. Furlong [25], M. Rautschka [26], M. L. Neves [27], A. Pascual-Montano [28], Q. Wei [29], N. Collier [30], M. F. M. Chowdhury [31], A. Lavelli [32], R. Berlanga [33], R. Morante [34], V. V. Asch [35], W. Daelemans [36], J. L. Marina [37], E. van Mulligen [38], J. Kors [39], and U. Hahn [40]
Journal	Journal of biomedical semantics
Volume	2 Suppl 5
Pagination	S11
Date Published	2011
ISSN	2041-1480
Abstract	BACKGROUND: Competitions in text mining have been used to measure the performance of automatic text processing solutions against a manually annotated gold standard corpus (GSC). The preparation of the GSC is time-consuming and costly and the final corpus consists at the most of a few thousand documents annotated with a limited set of semantic groups. To overcome these shortcomings, the CALBC project partners (PPs) have produced a large-scale annotated biomedical corpus with four different semantic groups through the harmonisation of annotations from automatic text mining solutions, the first version of the Silver Standard Corpus (SSC-I). The four semantic groups are chemical entities and drugs (CHED), genes and proteins (PRGE), diseases and disorders (DISO) and species (SPE). This corpus has been used for the First CALBC Challenge asking the participants to annotate the corpus with their text processing solutions. RESULTS: All four PPs from the CALBC project and in addition, 12 challenge participants (CPs) contributed annotated data sets for an evaluation against the SSC-I. CPs could ignore the training data and deliver the annotations from their genuine annotation system, or could train a machine-learning approach on the provided pre-annotated data. In general, the performances of the annotation solutions were lower for entities from the

categories CHED and PRGE in comparison to the identification of entities categorized as DISO and SPE. The best performance over all semantic groups were achieved from two annotation solutions that have been trained on the SSC-I. The data sets from participants were used to generate the harmonised Silver Standard Corpus II (SSC-II), if the participant did not make use of the annotated data set from the SSC-I for training purposes. The performances of the participants' solutions were again measured against the SSC-II. The performances of the annotation solutions showed again better results for DISO and SPE in comparison to CHED and PRGE.

CONCLUSIONS: The SSC-I delivers a large set of annotations (1,121,705) for a large number of documents (100,000 Medline abstracts). The annotations cover four different semantic groups and are sufficiently homogeneous to be reproduced with a trained classifier leading to an average F-measure of 85%. Benchmarking the annotation solutions against the SSC-II leads to better performance for the CPs' annotation solutions in comparison to the SSC-I.

URL

<http://www.jbiomedsem.com/content/2/S5/S11/> [41]

DOI

[10.1186/2041-1480-2-S5-S11](https://doi.org/10.1186/2041-1480-2-S5-S11) [42]

Copyright

© 2011 Rebholz-Schuhmann et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0> [43]), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Impact Factor

2.54 (Unofficial)

History

Published: 6 October 2011

Alternate Journal

J Biomed Semantics

Attachment

Size

[2041-1480-2-S5-S11.pdf](#) [44]

647.11 KB



Except where otherwise noted, all original content on this site is copyright by its author and licensed under a [Creative Commons Attribution-Share Alike 2.5 Canada License](#).

Source URL (retrieved on 2026-01-07 20:38):

<https://www.semanticsoftware.info/biblio/assessment-ner-solutions-against-first-and-second-calbc-silver-standard-corpus-0>

Links:

[1] <https://www.semanticsoftware.info/users/rene>

[2] <https://www.semanticsoftware.info/category/topic/bioinformatics>

[3] <https://www.semanticsoftware.info/category/topic/text-mining>

[4] <https://www.semanticsoftware.info/biblio/author/65>

[5] <https://www.semanticsoftware.info/biblio/author/136>

- [6] <https://www.semanticsoftware.info/biblio/author/87>
- [7] <https://www.semanticsoftware.info/biblio/author/88>
- [8] <https://www.semanticsoftware.info/biblio/author/89>
- [9] <https://www.semanticsoftware.info/biblio/author/90>
- [10] <https://www.semanticsoftware.info/biblio/author/91>
- [11] <https://www.semanticsoftware.info/biblio/author/92>
- [12] <https://www.semanticsoftware.info/biblio/author/26>
- [13] <https://www.semanticsoftware.info/biblio/author/25>
- [14] <https://www.semanticsoftware.info/biblio/author/93>
- [15] <https://www.semanticsoftware.info/biblio/author/14>
- [16] <https://www.semanticsoftware.info/biblio/author/1>
- [17] <https://www.semanticsoftware.info/biblio/author/94>
- [18] <https://www.semanticsoftware.info/biblio/author/16>
- [19] <https://www.semanticsoftware.info/biblio/author/137>
- [20] <https://www.semanticsoftware.info/biblio/author/138>
- [21] <https://www.semanticsoftware.info/biblio/author/97>
- [22] <https://www.semanticsoftware.info/biblio/author/98>
- [23] <https://www.semanticsoftware.info/biblio/author/99>
- [24] <https://www.semanticsoftware.info/biblio/author/100>
- [25] <https://www.semanticsoftware.info/biblio/author/139>
- [26] <https://www.semanticsoftware.info/biblio/author/102>
- [27] <https://www.semanticsoftware.info/biblio/author/140>
- [28] <https://www.semanticsoftware.info/biblio/author/141>
- [29] <https://www.semanticsoftware.info/biblio/author/105>
- [30] <https://www.semanticsoftware.info/biblio/author/106>
- [31] <https://www.semanticsoftware.info/biblio/author/142>
- [32] <https://www.semanticsoftware.info/biblio/author/108>
- [33] <https://www.semanticsoftware.info/biblio/author/143>
- [34] <https://www.semanticsoftware.info/biblio/author/110>
- [35] <https://www.semanticsoftware.info/biblio/author/111>
- [36] <https://www.semanticsoftware.info/biblio/author/112>
- [37] <https://www.semanticsoftware.info/biblio/author/116>
- [38] <https://www.semanticsoftware.info/biblio/author/144>
- [39] <https://www.semanticsoftware.info/biblio/author/145>
- [40] <https://www.semanticsoftware.info/biblio/author/27>
- [41] <http://www.jbiomedsem.com/content/2/S5/S11/>
- [42] <http://dx.doi.org/10.1186/2041-1480-2-S5-S11>
- [43] <http://creativecommons.org/licenses/by/2.0>
- [44] <https://www.semanticsoftware.info/system/files/2041-1480-2-S5-S11.pdf>