

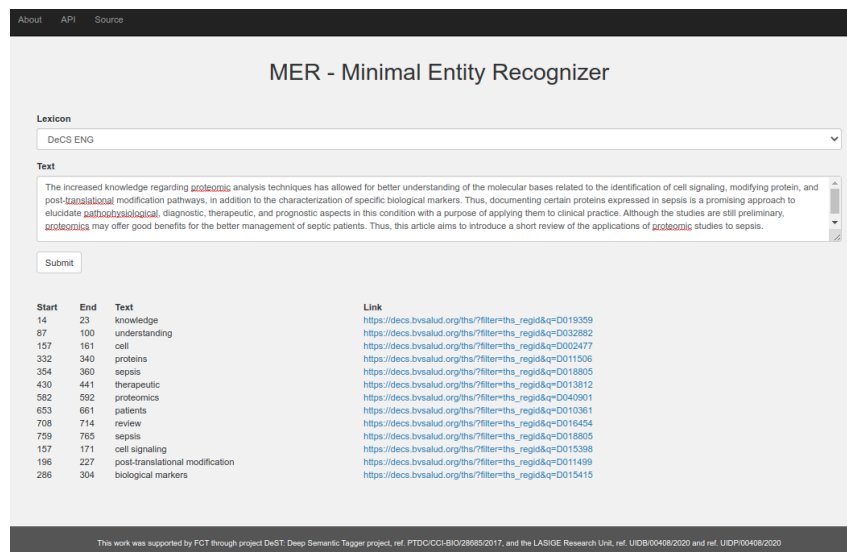
DeST: Deep Semantic Tagger

Francisco M. Couto ¹

¹ LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Besides the exponential growth of knowledge bases (KB) and the initiatives to connect them, most of our biomedical knowledge is still locked in free text. Free text has been and continues to be for humans the traditional and natural mean of representing and sharing knowledge. However, the knowledge encoded in free text hinders its accessibility and usage, since the retrieval of information from a large corpus is a tedious and time-consuming task for humans and a hard and prone to error task for machines. The ability to automatically process text enable us to more effectively navigate, retrieve information, find evidence, updates or even discern relevant from irrelevant information. Effectively linking text to KBs will also enhance the computer's ability to infer new knowledge. However, all these benefits require in-depth text mining solutions accessible to any researcher. With this goal in mind, the Deep Semantic Tagger (DeST) project developed open source software to perform the common tasks of a text mining pipeline, namely Named-Entity Recognition and Named-Entity Linking, Relation Extraction, and Semantic Similarity.

The performance of this software was extensively assessed in shared tasks using available corpora and datasets, with many of them achieving top-ranking positions in international challenges. Some of them are also available as web tools, such as the MER - Minimal Entity Recognizer (screenshot shown in the image). Given that most of these tools are based in machine learning techniques that require training data, the project also produced open access corpora and datasets that any researcher can use to enhance their models.



Start	End	Text	Link
14	23	knowledge	https://decs.bvsalud.org/ths/?filter=this_regid&q=D019359
87	100	understanding	https://decs.bvsalud.org/ths/?filter=this_regid&q=D032882
157	161	cell	https://decs.bvsalud.org/ths/?filter=this_regid&q=D002477
332	340	proteins	https://decs.bvsalud.org/ths/?filter=this_regid&q=D011506
354	360	sepsis	https://decs.bvsalud.org/ths/?filter=this_regid&q=D018805
430	441	therapeutic	https://decs.bvsalud.org/ths/?filter=this_regid&q=D013812
582	592	proteomics	https://decs.bvsalud.org/ths/?filter=this_regid&q=D049501
653	661	patients	https://decs.bvsalud.org/ths/?filter=this_regid&q=D010381
708	714	review	https://decs.bvsalud.org/ths/?filter=this_regid&q=D016454
759	765	sepsis	https://decs.bvsalud.org/ths/?filter=this_regid&q=D018805
157	171	cell signaling	https://decs.bvsalud.org/ths/?filter=this_regid&q=D015398
196	227	post-translational modification	https://decs.bvsalud.org/ths/?filter=this_regid&q=D011499
286	304	biological markers	https://decs.bvsalud.org/ths/?filter=this_regid&q=D015415

The project also published an open access book[1] that aims at helping Health and Life specialists or students to easily learn how to process data and text, by showing how shell scripting can help solve many of the data processing tasks that Health and Life specialists face everyday with minimal software dependencies.

The open source software, corpora and datasets are available at: <https://github.com/lasigeBioTM>; the web tools and all the book material at <http://labs.rd.ciencias.ulisboa.pt/>

The open source software, corpora and datasets are available at: <https://github.com/lasigeBioTM>; the web tools and all the book material at <http://labs.rd.ciencias.ulisboa.pt/>

[1] Couto, F. M. (2019). Data and text processing for health and life sciences (p. 98). Springer Nature.

Acknowledgements

This work was supported by FCT through project DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017, and the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020"