Preface of the First International Biochemical Knowledge Extraction Challenge (BiKE)

Edgard Marx[1], Marilia Valli[2], Joao da Silva e Silva[2], Sanju Tiwari[3] and Paulo do Carmo[1]

[1]*Leipzig University of Applied Science, Germany*
[2]*Sao Paulo University, Brazil*
[3]*Universidad Autonoma de Tamaulipas, Mexico*

The knowledge of over 50 years of studies on biodiversity available in scientific articles can become easier accessible when organized and shared through knowledge graphs. It can assist in the development of different fields of science and bio-friendly products with high added value as well as guide public policies to bring benefits both to science and to strengthen the bio-economy. However, to date, most of the structured biochemical information available on the Web is manually curated, and it is practically impossible to keep pace with the research being constantly published in scientific articles.

The First International Biochemical Knowledge Extraction Challenge (BiKE) aims at accelerating and promoting the research on automatic biochemical knowledge extraction mechanisms by the Semantic Web scientific community to increase the information available on natural products and contribute to the development of environmental-friendly products while increasing the community awareness of the biodiversity value. The following papers were accepted for publication and presented at the workshop:

- BiKE Challenge: Result of ChemiScope by using ChatGPT
- Improving Natural Product Automatic Extraction with Named Entity Recognition
- Enhancing Biochemical Extraction with BFS-driven Knowledge Graph Embedding approach

Challenge

BiKE challenge invited researchers to participate by re-using or designing new innovative Biochemical Extraction methods. The challenge consisted of extracting relevant information from biochemical research articles and constructing a Biochemical Knowledge Graph (BKG) through a given ontology. Biochemical Knowledge Graphs are knowledge graphs containing bio- and chemical information from living organisms.

BiKE 2023: First International Biochemical Knowledge Extraction Challenge, Co-located with the ESWC 2023, May 05-29-2023, Crete, Hersonissos, Greece
edgard.marx@htwk-leipzig.de (E. Marx); marilia.valli@ifsc.usp.br (M. Valli); jvictor.ssilva@ifsc.usp.br (J. d. S. e. Silva); tiwarisanju18@ieee.org (S. Tiwari); paulo.carmo@htwk-leipzig.de (P. d. Carmo)

Training & Test Data sets

The dataset used for evaluation and training was generated from hundreds of peer-reviewed scientific articles with information on more than 2,521 possibilities of natural product extraction. The dataset was built manually by chemistry specialists that read the articles annotating four relevant properties associated with each natural product discussed in the academic publication. For this challenge, we focus on five NuBBE properties for training and prediction: (I) compound name (rdfs:label), (II) bioactivity (nubbe:biologicalActivity), (III) species from where natural products were extracted (nubbe:collectionSpecie), (IV) collection site of these species (nubbe:collectionSite), and (V) isolation type (nubbe:collectionType). The table below presents an overview of the number of unique properties.

All papers are present on all train splits, but the papers selected for each test split have all links to manually extracted characteristics removed. This means that these papers were not connected to the rest of the knowledge graph. The provided code for generating the knowledge graph representation with Python's networkx uses BERTopic's extracted topics for reconnecting the knowledge graph. The assigned topics were also filtered by the following rule: if the topic is present in more than 80% of examples, it is eliminated since it does not discriminate from the others. Part of the challenge was to figure out other ways to reconnect the knowledge graph with automatically extracted characteristics like citation networks for the authors, conferences, and others.

The challenge provided the original flat data and the original networkx knowledge graph. It also provided 10 previously randomized train/test splits that contain the links maintained and removed, respectively. For every train/test split, we also provide a prepared networkx knowledge graph. The source code and documentation for the benchmark dubbed as NatUKE [1] is publicly available at https://github.com/AKSW/natuke.

Evaluation Metrics

The challenge was focused on ranking the correct document prediction of real links that were hidden in the knowledge graph. Together with MRR (Mean Reciprocal Rank), hits@k is a ranking metric for when there is only one correct document. On the other hand, mAP (mean Average Precision) and nDCG (normalized Discounted Cumulative Gain) are designed for ranking when a list of relevant documents is available. The hits@k was chosen because it allows the evaluation of each characteristic extraction with reasonable expectations by customizing the k value. Following the rule used in NatUKE the final k values in this table are from 1 to 50 considering values multiples of 5 and two thresholds: (1) a score equal or higher than 0.50 is achieved; and (2) a score equal or higher than 0.20 is achieved. Please refer to the NatUKE benchmark paper for further details.

Talks

- **Towards Natural Inspired Products from Biodiversity**
  Edgard Marx, Leipzig University of Applied Sciences (HTWK), Germany

- **NaTUKE: A Benchmark for Natural Product Knowledge Extraction from Academic Literature**
  Paulo Ricardo Viviurka do Carmo, Leipzig University of Applied Sciences (HTWK), Germany
- **The NuBBE Knowledge Graph: A Biochemical Knowledge Graph of Natural Products from Brazilian Biodiversity**
  István J. Mócsy, Leipzig University of Applied Sciences (HTWK), Germany

Best Knowledge Extraction Awards

The Best Extraction Method award was intended to recognize the top three competitors who have demonstrated exceptional abilities, commitment, and a comprehensive comprehension of the ideas and procedures involved in extracting pertinent data from challenging biochemical datasets. These people have demonstrated outstanding critical thinking, problem-solving skills, and a thorough understanding of cutting-edge computational tools and methods. The prizes were given with a unique certificate of appreciation that features the winner's name and special workshop accomplishments, followed by a monetary reward. The three winners of the first edition of the BiKE challenge were:

1st **BiKE Challenge: Result of ChemiScope by using ChatGPT**
Matthias Jooß, Jonas Gwozdz and Pit Fröhlich

2nd **Improving Natural Product Automatic Extraction with Named Entity Recognition**
Stefan Schmidt-Dichte and István J. Mócsy

3rd **Enhancing Biochemical Extraction with BFS-driven Knowledge Graph Embedding approach**
Bhushan Zope, Sashikala Mishra and Sanju Tiwari

General Chair

- Edgard Marx, Leipzig University of Applied Sciences (HTWK), Germany

Organizing Committee

- Marilia Valli, Sao Paulo University (USP), Brazil
- Joao Victor da Silva e Silva, Sao Paulo University (USP), Brazil
- Sanju Tiwari, Universidad Autonoma de Tamaulipas (UAT), Mexico
- Paulo Ricardo Viviurka do Carmo, Leipzig University of Applied Sciences (HTWK), Germany

Advisory Committee

- Vanderlan da Silva Bozani, Sao Paulo State University (UNESP), Brazil
- Adriano Defini Andricopulo, Sao Paulo University (USP), Brazil
- Thomas Riechert, Leipzig University of Applied Sciences (HTWK), Germany
- Alan Pilon, Sao Paulo University (USP), Brazil

Acknowledgements

The editors would like to thank the advisory team, authors, program committee, and other organizers for their constant support in making this event successful.

References

[1] P. V. Do Carmo, E. Marx, R. Marcacini, M. Valli, J. V. Silva e Silva, A. Pilon, NatUKE: A Benchmark for Natural Product Knowledge Extraction from Academic Literature, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), 2023, pp. 199–203. doi:10.1109/ICSC56153.2023.00039.