Constructing Information-Lossless Biological Knowledge Graphs from Conditional Statements

Tianwen Jiang[1,2], Tong Zhao[1], Bing Qin[2], Ting Liu[2], Nitesh V. Chawla[1], Meng Jiang[1]

[1]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

[2]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

twjiang@ir.hit.edu.cn,tzhao2@nd.edu,{bqin,tliu}@ir.hit.edu.cn,{nchawla,mjiang2}@nd.edu

ABSTRACT

Conditions are essential in the statements of biological literature. Without the conditions (e.g., environment, equipment) that were precisely specified, the facts (e.g., observations) in the statements may no longer be valid. One biological statement has one or multiple fact(s) and/or condition(s). Their subject and object can be either a concept or a concept's attribute. Existing information extraction methods do not consider the role of condition in the biological statement nor the role of attribute in the subject/object. In this work, we design a new tag schema and propose a deep sequence tagging framework to structure conditional statement into fact and condition tuples from biological text. Experiments demonstrate that our method yields a information-lossless structure of the literature.

1 INTRODUCTION

Extracting information from biological text plays an important role in biological knowledge graph construction, relational inference, hypothesis generation and validation. Open IE systems extract a diverse set of relational tuples without requiring any relation-specific schema in advance, which was supposed to be ideally suited to the biological corpora. For example, given a statement sentence in a biochemistry paper [3]: "*We observed that ... alkaline pH increases the activity of TRPV5/V6 channels in Jurkat T cells.*", an Open IE system [2] would return a (subject, relation, object)-tuple (*alkaline pH, increases, activity of TRPV5/V6 channels in Jurkat T cells*).

Such an information representation has two problems. First, the condition that the channels were in the Jurkat T cells, on which the observation was obtained, remained unstructured in the object argument. In biological domains, conditions are essential in claiming observations and hypotheses [1]. Second, the object name was long, infrequent, and less likely to be associated with other tuple extractions, leading to knowledge graph sparsity and poor inference, though the concept "TRPV5/V6 channels" and its attribute "activity" could be frequently mentioned. Unlike the general domain where relations are often linking two named entities, in biological domains, the relations could be linking either biological concepts or a concept's particular attributes. A well-designed information structure of biological statement is desired.

We propose a new information extraction task named "Biological Conditional Statement Structuring" (BioCS). Given a biological statement sentence, BioCS extracts fact tuple(s) as well as condition tuple(s) on which the fact(s) were observed or claimed. The expected outputs of the example sentence are: **Fact 1:** (alkaline_pH, increases, {TRPV5/V6 _channels: activity}), and **Condition 1:** (TRPV5/V6_channels, in, Jurkat _T_cells). The subject or object in the tuple is formatted as {concept: attribute} where the attribute could be null if it is a concept only. Actually we find
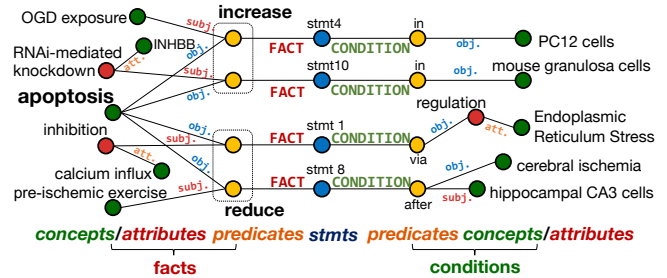

Figure 1: Our biological knowledge graph provides a visualized, comprehensive understanding of what increased/reduced "apoptosis" under what kind of conditions.

another fact tuple from the full example sentence: **Fact 2:** (extracellular_acidic_pH, reduces, {TRPV5/V6_channels: activity})

Inspired by [2], we formulate BioCS as a sequence tagging problem. We propose a new deep sequence tagging framework to achieve the goal. Experiments on a data set of 141M sentences from PubMed paper abstracts show that the biological knowledge graph we constructed provide a good understanding of biological statements (https://scikg.github.io).

2 FRAMEWORK DESCRIPTION

Our framework has two modules: (1) a deep sequence tagging model, taking multiple language feature sequences as inputs and returning multiple fact and condition tuples; (2) an iterative self-training scheme with massive unlabeled data to enhance the model.

3 PRELIMINARY RESULTS

Suppose we are interested in what *increased*/*decreased* "apoptosis". The KG provides us a snapshot in Figure 1: (1) "OGD exposure" and the "RNAi-mediated knockdown" of "INHBB" increased apoptosis, and (2) the "inhibition" of "calcium influx" and "pre-ischemic exercise" reduced apoptosis with the left side of the figure. It is important to be aware of the condition for each factual claim on the right side of the figure. They describe either the methodology of the observation or the context of it. This KG will enable effective biological knowledge inference and reasoning.

REFERENCES

[1] David L Miller. 1947. The Nature of Scientific Statements. *Philosophy of Science* 14, 3 (1947), 219–223.

[2] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *NAACL*. 885–895.

[3] Victor N Tomilin, Alena L Cherezova, Yuri A Negulyaev, and Svetlana B Semenova. 2016. TRPV5/V6 channels mediate Ca2+ influx in jurkat T cells under the control of extracellular pH. *Journal of cellular biochemistry* 117, 1 (2016), 197–206.