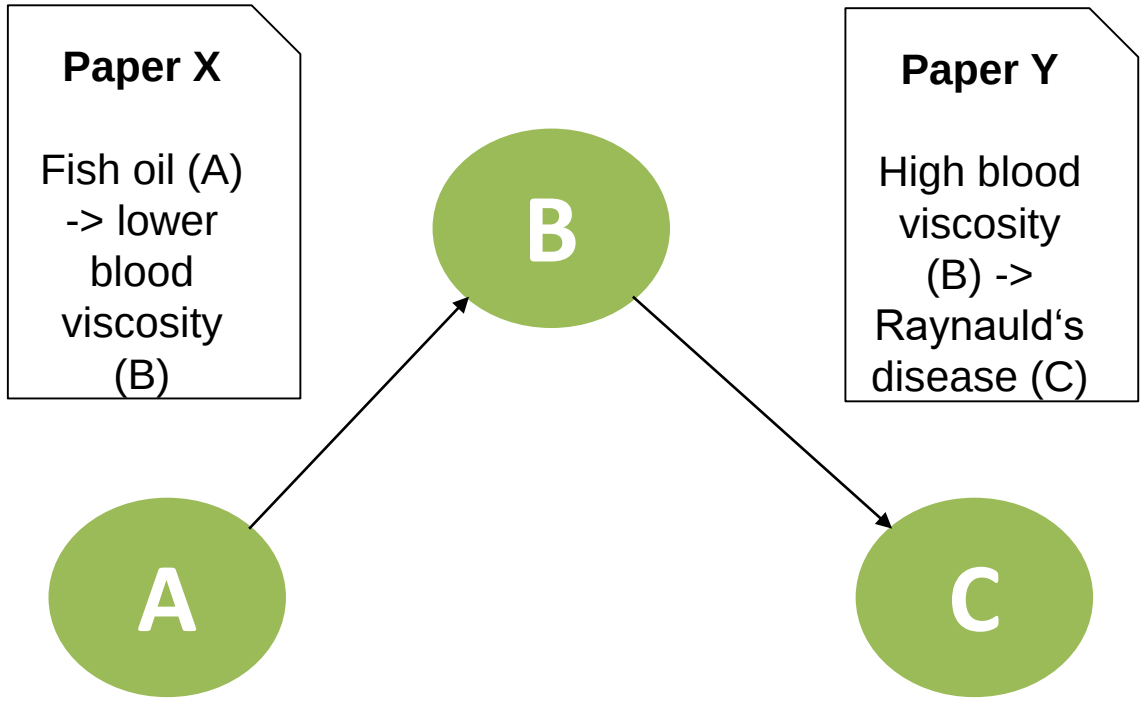
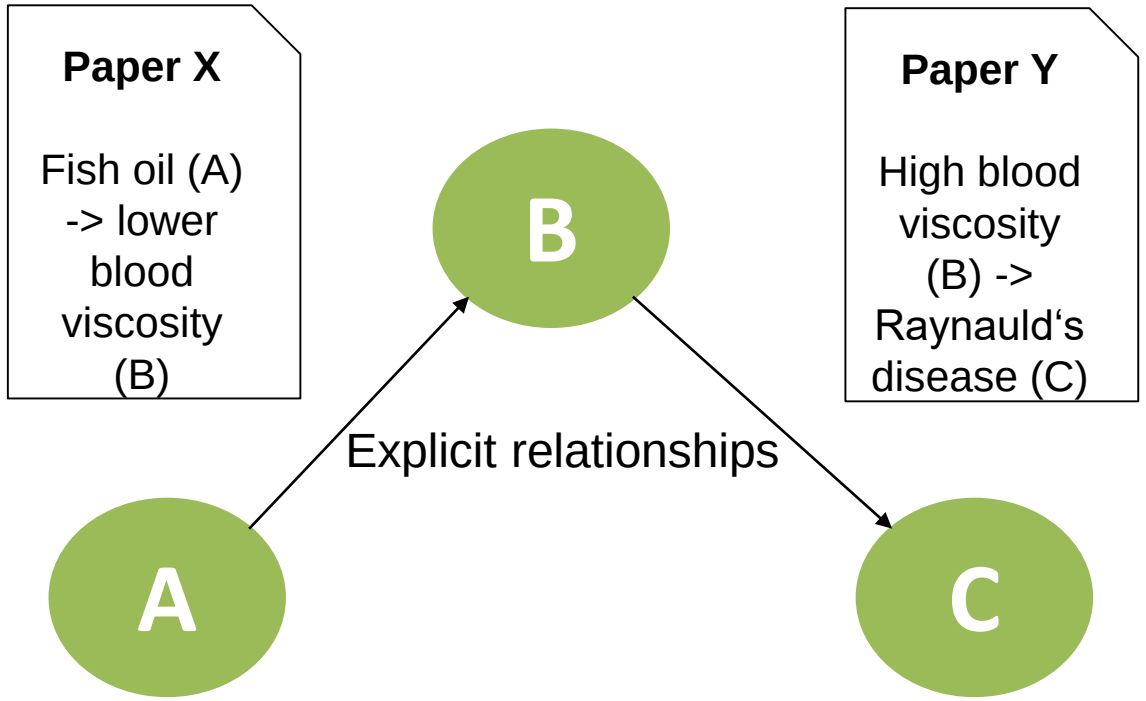
# Extracting and Connecting Scientific Knowledge from Texts

Tobias Hecking
*Institute for Software Technology – Dept. for Intelligent and Distributed Systems*

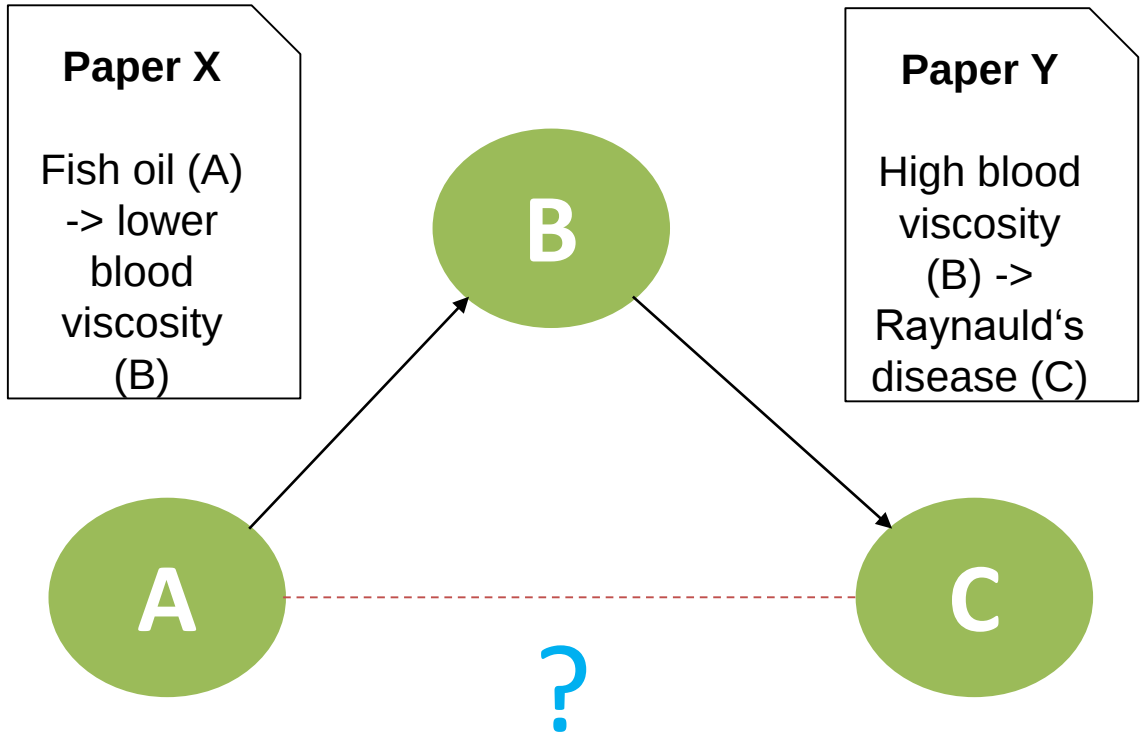
# Motivation: Discovering implicit knowledge in publication databases



# Motivation: Discovering implicit knowledge in publication databases



# Motivation: Discovering implicit knowledge in publication databases



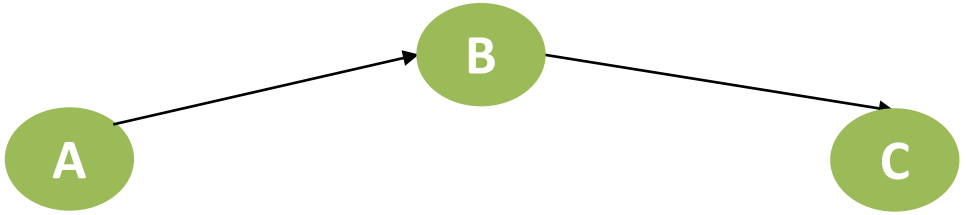

Implicit knowledge: Fish oil has an effect on Raynauld‘ desease

Swanson, Don R. "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge." *Perspectives in Biology and Medicine*, vol. 30 no. 1, 1986, p. 7-18. *Project MUSE*, doi:10.1353/pbm.1986.0087.

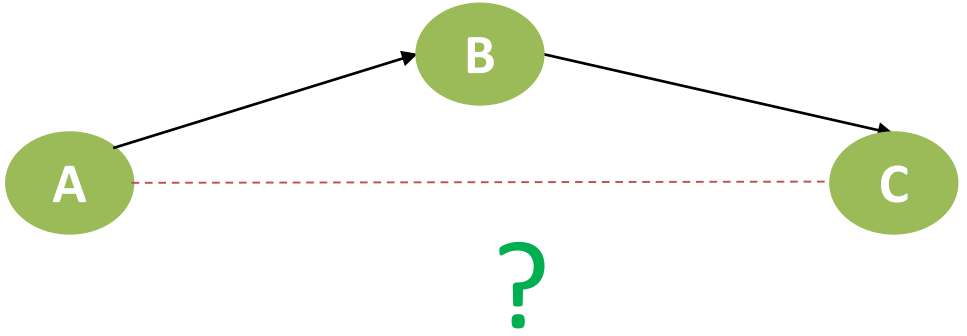
# Scientific knowledge graph extraction workflow

1) Detect important concepts in texts


2) Relation identification




3) Analysis and querying

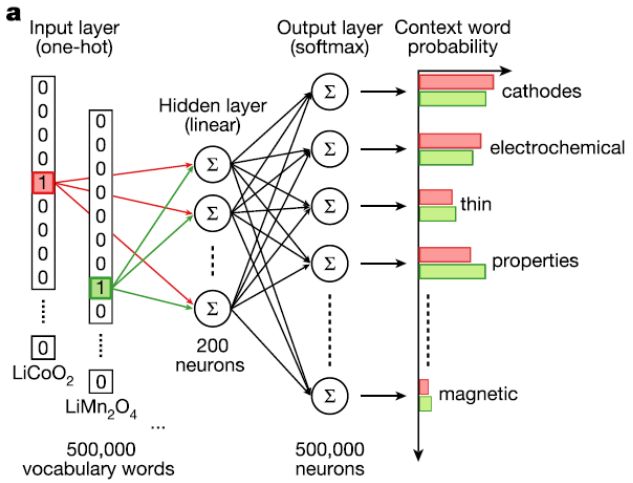


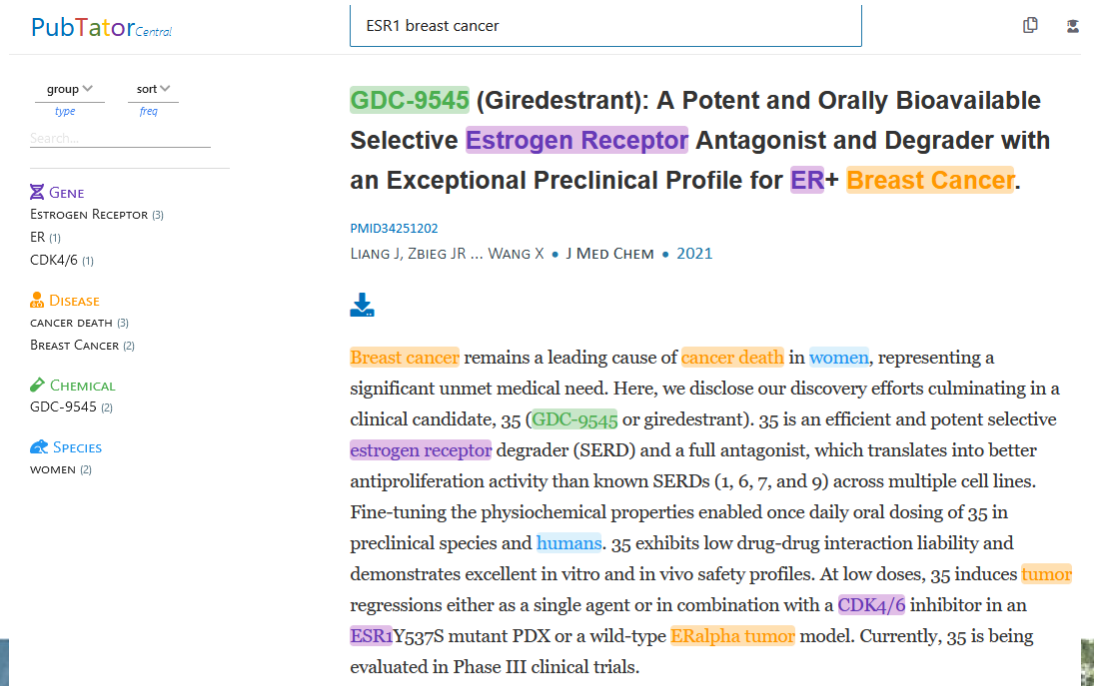
# Many language resources for biomedicine and chemistry but only a few for other domains

- UMLS
- Lion LBD
- PubTator

- SciBERT (SciVocab)
- STEM-ECR
- SciERC






DLR

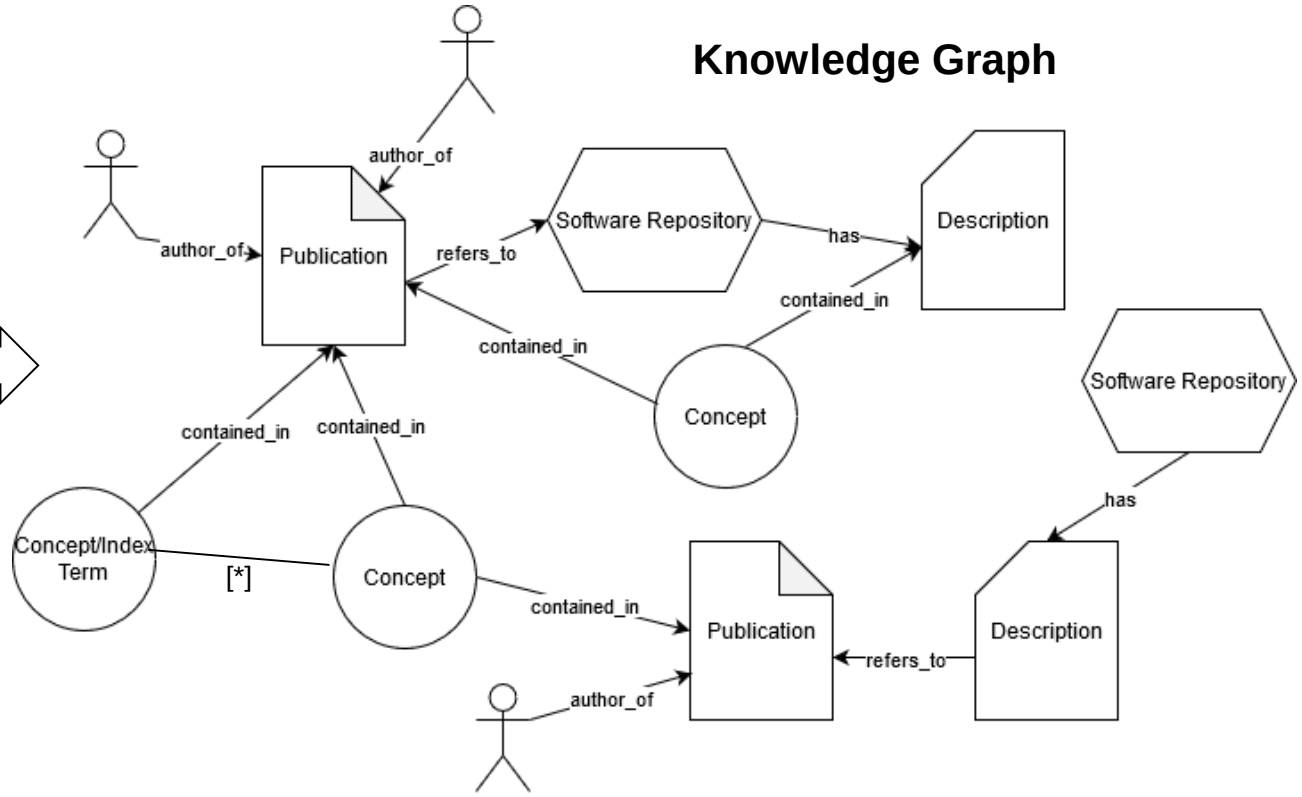
DLR

# Towards open domain scientific knowledge graph extraction

GitLab

elib
Publikationen des DLR

**Crawling + Information Extraction**




- Goal: Enable complex search queries:
  - Find data and code for a given paper?
  - Expert finding?
  - Trend detection and literature-based discovery.

# Concept Identification

## (1) Wikipedia as knowledge base

Intelligent engine control could be one of the most important innovations in the development of future reusable engines, facilitating a safer and more economical engine operation. In this work, we investigate the closed-loop control of the LUMEN expander-bleed engine by combining machine learning with a transient simulation environment. The controller can dynamically change the set-point of the engine between a chamber pressure of 40 bar to 80 bar by adjusting up to six flow control valves while maintaining several boundary conditions at any given time.

# Concept Identification

## (1) Wikipedia as knowledge base

Intelligent engine control could be one of the most important innovations in the development of future reusable engines, facilitating a safer and more economical engine operation. In this work, we investigate the closed-loop control of the LUMEN expander-bleed engine by combining machine learning with a transient simulation environment. The controller can dynamically change the set-point of the engine between a chamber pressure of 40 bar to 80 bar by adjusting up to six flow control valves while maintaining several boundary conditions at any given time.

## (2) Named Entity Recognition*

Intelligent engine control could be one of the most important innovations in the development of future reusable engines, facilitating a safer and more economical engine operation. In this work, we investigate the closed-loop control of the LUMEN expander-bleed engine by combining machine learning with a transient simulation environment. The controller can dynamically change the set-point of the engine between a chamber pressure of 40 bar to 80 bar by adjusting up to six flow control valves while maintaining several boundary conditions at any given time.

*https://spacy.io/

# Usage of Text-based QA Systems

## (1) Wikipedia als Wissensbasis

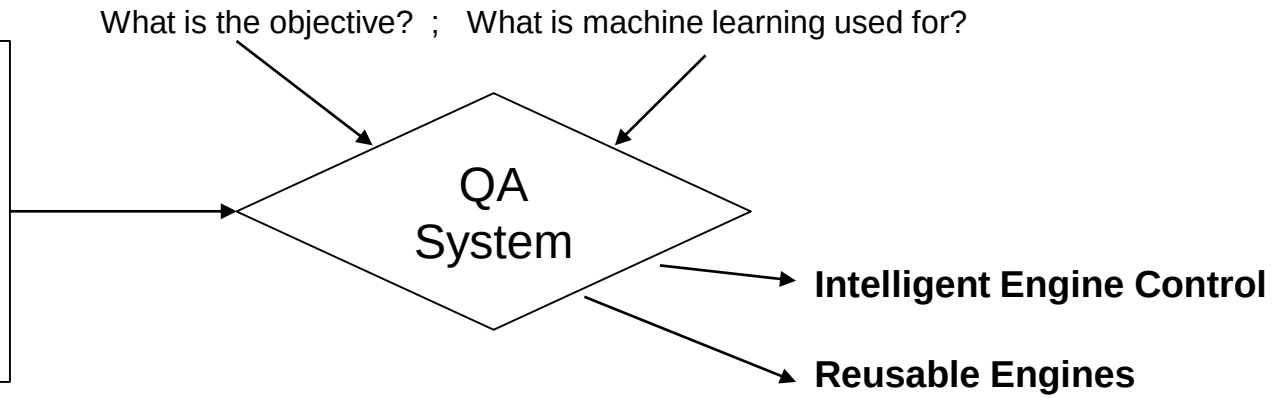
Intelligent engine control could be one of the most important innovations in the development of future reusable engines, facilitating a safer and more economical engine operation. In this work, we investigate the closed-loop control of the LUMEN expander-bleed engine by combining machine learning with a transient simulation environment. The controller can dynamically change the set-point of the engine between a chamber pressure of 40 bar to 80 bar by adjusting up to six flow control valves while maintaining several boundary conditions at any given time.

## (2) Named Entity Recognition

Intelligent engine control could be one of the most important innovations in the development of future reusable engines, facilitating a safer and more economical engine operation. In this work, we investigate the closed-loop control of the LUMEN expander-bleed engine by combining machine learning with a transient simulation environment. The controller can dynamically change the set-point of the engine between a chamber pressure of 40 bar to 80 bar by adjusting up to six flow control valves while maintaining several boundary conditions at any given time.

## (3) Usage of Text-based QA Systems*

Intelligent engine control could be one of the most important innovations in the development of future reusable engines, facilitating a safer and more economical engine operation. In this work, we investigate the closed-loop control of the LUMEN expander-bleed engine by combining machine learning with a transient simulation environment. The controller can dynamically change the set-point of the engine between a chamber pressure of 40 bar to 80 bar by adjusting up to six flow control valves while maintaining several boundary conditions at any given time.




*https://arxiv.org/pdf/2011.00057.pdf

# QA for NER

| Entity | Question |
|---|---|
| Process | "What is the process?" |
| Method | "What is the method?" |
| Material | "Which material is used?" |
| Data | "Which data is used?" |

Table 3.1: QA for NER on STEM-ECR, entity question mapping.

# Information Linking Example

## Publication abstract

Intelligent engine control could be one of the most important innovations in the development of future reusable engines, facilitating a safer and more economical engine operation. In this work, we investigate the closed-loop control of the LUMEN expander-bleed engine by combining machine learning with a transient simulation environment. The controller can dynamically change the set-point of the engine between a chamber pressure of 40 bar to 80 bar by adjusting up to six flow control valves while maintaining several boundary conditions at any given time.

Intelligent Engine Control

Reusable Engines

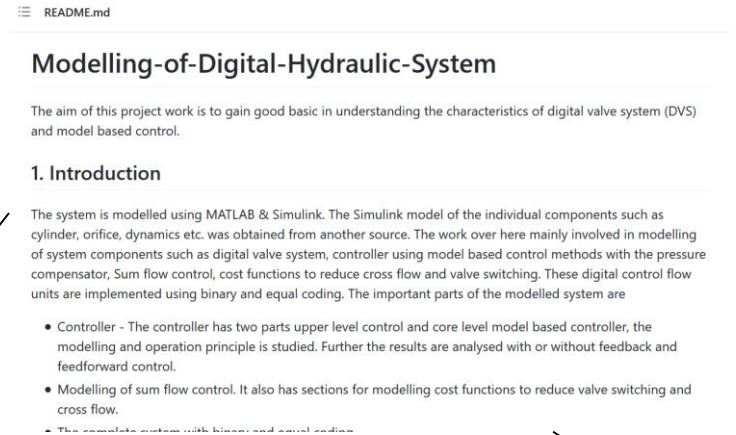
Machine Learning

LUMEN

Simulation Environment

Flow Control Valves

## GitHub Project READ.ME file


README.md

Modelling-of-Digital-Hydraulic-System

The aim of this project work is to gain good basic in understanding the characteristics of digital valve system (DVS) and model based control.

1. Introduction

The system is modelled using MATLAB & Simulink. The Simulink model of the individual components such as cylinder, orifice, dynamics etc. was obtained from another source. The work over here mainly involved in modelling of system components such as digital valve system, controller using model based control methods with the pressure compensator, Sum flow control, cost functions to reduce cross flow and valve switching. These digital control flow units are implemented using binary and equal coding. The important parts of the modelled system are

- Controller - The controller has two parts upper level control and core level model based controller, the modelling and operation principle is studied. Further the results are analysed with or without feedback and feedforward control.
- Modelling of sum flow control. It also has sections for modelling cost functions to reduce valve switching and cross flow.

Simulink

Model Based Control

# Ongoing work: Neural methods for concept extraction and classification

- Use of SciBERT (Special BERT language model trained on scientific texts)
- Train model on annotatated STEM-ECR dataset

  Annotations: Process, Method, Material, Data

| Model | precision | recall | F1 | Accuracy |
|---|---|---|---|---|
| bert-large-uncased | 62,6 | 70,1 | 66,17 | 86,6 |
| **SciBERT (paper)** | 64,3 | 66,7 | 65,77 | ? |
| bert-large-cased | 63,87 | 69,2 | 66,4 | 86,6 |
| roberta-large | 62,9 | 68,5 | 65,6 | 86,5 |
| google/electra-base-discriminator | 64,31 | 66,7 | 65,5 | 86,8 |
| allenai/scibert-scivocab-cased | 61,9 | 69,2 | 65,4 | 86,0 |
| roberta-base | 62,9 | 67,1 | 65,0 | 85,7 |
| allenai/scibert-scivocab-uncased | 60,7 | 68,8 | 64,5 | 86,7 |
| google/electra-large-discriminator | 62,1 | 66,4 | 64,2 | 85,7 |
| bert-base-cased | 58,0 | 66,6 | 62,0 | 85,9 |
| bert-base-uncased | 58.0 | 65,4 | 61,5 | 84,9 |

# Ongoing work: QA for Relation Extraction

| Relation | Question |
|---|---|
| USED-FOR | "What is E1 used for?" |
| FEATURE-OF | "What is E1 a feature of?" |
| CONJUNCTION | "What is E1 a conjunction of?" |
| HYPONYM-OF | "What is E1 a hyponym of?" |
| PART-OF | "What is E1 a part of?" |
| EVALUATE-FOR | "What is E1 evaluated for?" |
| COMPARE | "What is E1 compared with?" |

# Results: SciERC – QA

| Model | Basemodel | Correct |
|---|---|---|
| deepset/roberta-base-squad2 | roberta-base | 24.3 |
| deepset/xlm-roberta-large-squad2 | xlm-roberta-large | 18.4 |
| ktrapeznikov/albert-xlarge-v2-squad-v2 | albert-xlarge-v2 | 17.6 |
| deepset/electra-base-squad2 | google/electra-base-discriminator | 22.7 |
| distilbert-base-uncased-distilled-squad | distilbert-base-uncased | 14.79 |
| ahotrod/electra-large$_d$$iscriminator - squad2 - 512$ | google/electra-large-discriminator | 29.9 |

| Model | Correct relation | Correct entity |
|---|---|---|
| deepset/xlm-roberta-large-squad2 | 37.1 | 48.5 |
| ktrapeznikov/albert-xlarge-v2-squad-v2 | 32.8 | 47,6 |
| ahotrod/electra-large-discriminator-squad2-512 | 44.7 | 50,6 |



# Thank you!

tobias.hecking@dlr.de