Natural Language Processing for Nuclear Data

Nuclear Science & Security Consortium

Laura Shi[1], Bethany L. Goldblum[1,2], Walid Younes[2], Juan Manfredi[3], Jonathan Li[1], Char Juin Chin[1]

[1]Department of Nuclear Engineering, University of California, Berkeley, California 94720 USA
[2]Lawrence Berkeley National Laboratory, California 94720 USA
[3] Air Force Institute of Technology, Wright-Patterson AFB, Ohio 45433

Introduction

NucScholar is a web-based software framework in development that uses Natural Language Processing (NLP) to automatically retrieve, categorize, and recommend nuclear science papers. The goal of NucScholar is to provide the groundwork for a shift to a fully automated workflow for nuclear science literature searches, enabling increased efficiency in the nuclear data pipeline and accelerating data throughput for a wide range of applications. Please visit our website at nucscholar.berkeley.edu to learn more.

Background

- The current means of identifying and processing nuclear data bibliographic information is through the Nuclear Science References (NSR) database.
- NSR is heavily reliant on human intelligence tasks– evaluators manually identify articles and keywords.
- NLP techniques implemented in NucScholar can drastically improve the efficiency and effectiveness of the current workflow.

State-of-the-Art	NucScholar Innovation
Evaluators manually check over 80 major physics journals for relevant articles	Nuclear physics literature is automatically identified and collated with application programming interfaces (APIs), web crawling, and web scraping
Evaluators read a subset of articles and prepare database entries	All relevant articles are categorized as to their relevance to major physics topic areas; Database entries are generated using NLP techniques
Database entries are governed by a fixed set of predefined keywords	Adaptive database with emergent keywords that evolve with the evolving literature
Database queries require specific syntax and format	Natural language queries enable users to enter search terms in their own words

Table 1. *Comparison of the state-of-the-art (NSR) workflow with that of NucScholar*

Data Processing

- For training and validation, full NSR database encoded in human-readable and standardized JSON format.
- Metadata extraction via the CrossRef API enables NucScholar provides missing information (e.g., authors, titles, DOI, etc.) to complete current NSR entries *and to generate new entries.*
- Key nuclides are identified along with line number and frequency.

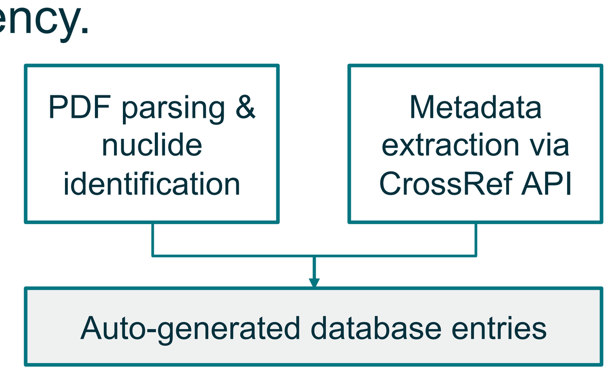

Fig 1. *Workflow for NucScholar library generation*

NucScholar Workflow

- NucScholar leverages recent developments in NLP techniques for named entity recognition, topic modeling, and sematic similarity using TextRank [1] and Latent Semantic Analysis (LSA) [2].
- Deep learning techniques with transformer models are currently being explored to augment existing NLP models and develop more complex systems like Question Answering (QA) models.

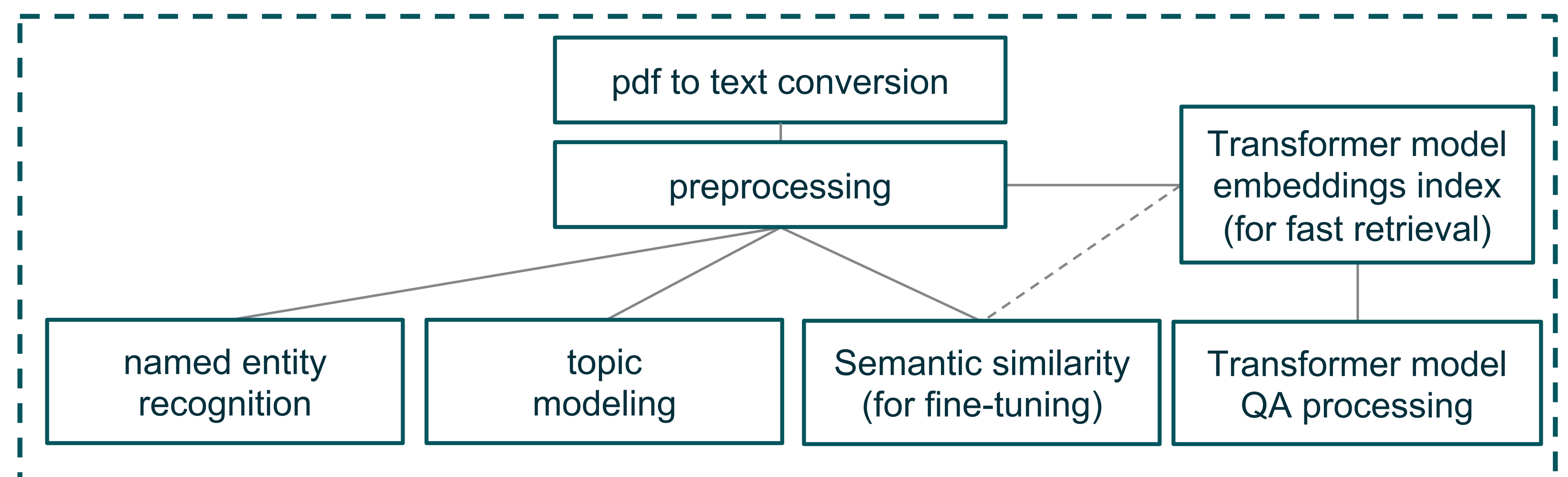

Fig 2. *Current workflow for NucScholar to support evaluation through natural language processing in the nuclear data pipeline. The user would interact through keyword and natural language queries*

- Topic modeling is performed using Latent Sematic Analysis to generate topic encoded data. For the training set, the topic vectors for papers from a given NSR subject area are averaged. For each paper in the test set, the topic vector is compared against the NSR subject topic vector using cosine similarity.

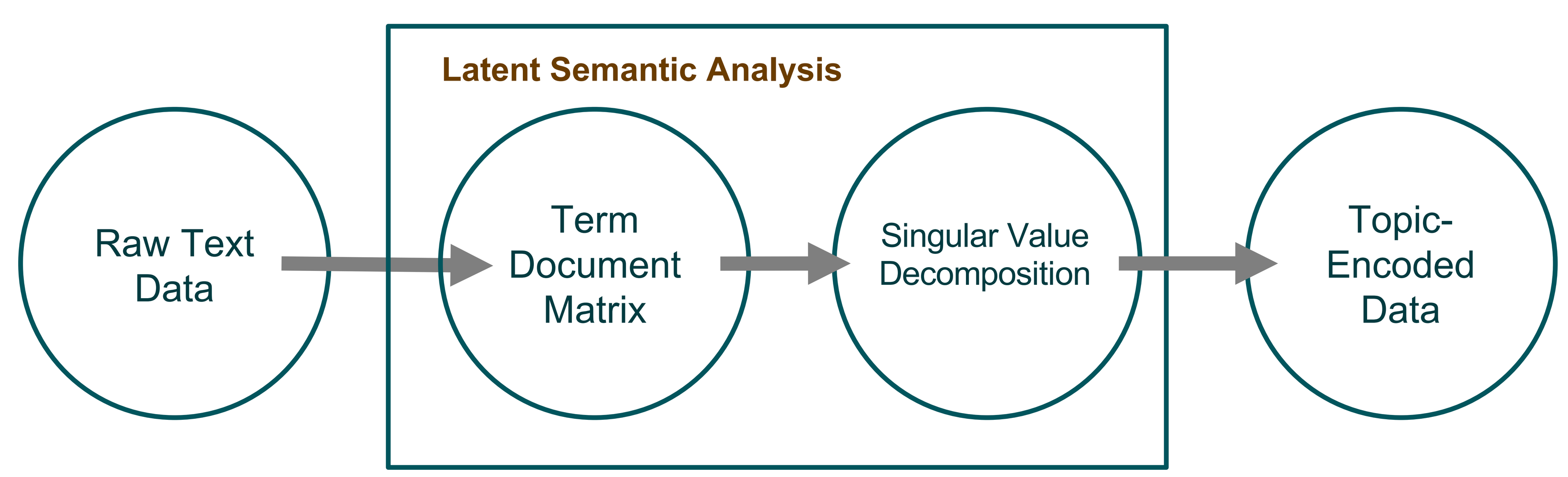

Fig 3. *Latent Sematic Analysis is used to classify literature into NSR subject area*

- A transformer is a deep learning model that weighs the influence of different parts of input data. The Deep Learning QA pipeline leverages the pre-trained BERT model and existing python modules to:
  1. Generate training data from nuclear science papers- pairs semantically similar sentences with a similarity score (custom NucScholar design)
  2. Fine-tune BERT using nuclear-specific data
  3. Generate and save sentence embeddings index
  4. Process user queries using embeddings index through semantic searches and extractive QA

```
Initializing BERT...
Fetching text
Enter question: how are data organized in xundl?
Q: how are data organized in xundl?
A: by nuclide
```

Fig 4. *Output from BERT model fine-tuned with nuclear data subject matter text*

Mission Relevance

- Applications of NucScholar pertinent to the nonproliferation research and development (NA-22) mission include [4]:
  - Identification of deficiencies in the U.S. Nuclear Data Program databases for national security applications.
  - Support of global nonproliferation norms with further development of analysis and computational capabilities to verify arms control and nonproliferation treaty commitments.
  - Recommendation of literature to drive work in special nuclear material (SNM) accounting, contraband detection, radiation shielding design, and advanced nuclear energy sector applications.

Fig 5. *Potential nuclear security and nonproliferation applications of NucScholar; a Non-Destructive Assay (NDA) methods setup for fuel assay and forensics (left) [5], NucScholar can be used to automate searches for online resources used with facility measurements for IAEA safeguards verification assessments (right) [6].*

References

1. Mihalcea, R. and Tarau, P (2004). TextRank: Bringing Order into Text. EMNLP, 4, 404-411.
2. Tshitoyan, V et al (2019). Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. Nature, 571, 95-98.
3. Sun, C et al (2020). How to Fine-Tune BERT for Text Classification. Fudan University.
4. Nuclear Data Needs and Capabilities for Applications (2015). Lawrence Berkeley National Laboratory.
5. Croft, S. and S.J. Tobin (2011). A Technical Review of Non-Destructive Assay Research for the Characterization of Spent Nuclear Fuel Assemblies Being Conducted Under the US DOE NGSI, LANL Report LA-UR-10-08045.
6. Green, A. and Dixit, A (2016). IAEA Safeguards Labs More Efficient and Accurate Thanks to Recent Upgrades.

Acknowledgements

This work was supported in part by the by the U.S. Department of Energy, National Nuclear Security Administration, Office of Defense Nuclear Nonproliferation Research and Development (DNN R&D) through the Nuclear Science and Security Consortium under Award Numbers DE-NA0003180 and DE-NA0003996.

This material is based upon work supported in part by the Department of Energy National Nuclear Security Administration through the Nuclear Science and Security Consortium under Award Number DE-NA0003996.