Ontology Learning System (OLS)

Sundos Al Subhi and **Chetan Tiwari** and **Armin R. Mikler**
Department of Computer Science
Georgia State University
Atlanta, GA, USA

Abstract

System for Ontology Learning and Extraction (SOLE) aims to automate hazard-specific ontology construction from knowledge bases of disaster-related information (e.g., scholarly articles) through the use of ontology learning techniques. The hazard-specific ontologies that are extracted from knowledge bases of disaster-related information will provide planners, policymakers, and decision-makers with the information they need in cases of disaster. This research will contribute by enabling the automated extraction and organization of unstructured data into structured data and information related to a crisis resulting from specific hazards. The proposed system, SOLE can be used to process real-time data from social media to uncover the effects of disasters in different locations, thereby improving critical disaster relief efforts. Also, this research will identify place and hazard-specific impacts by integrating formal and informal terms. Such information can provide critical intelligence for improving disaster planning, recovery, and resilience efforts. SOLE contains two components, which are the Ontology Learning System (OLS) and Semantic Mapping System (SMS). The extended abstract focuses only on the first component.

1 Introduction

In the United States, floods are the second deadliest of all-weather-related hazards with approximately 98 deaths per year (CDC, 2020). Planning for flooding events is critical to mitigate their impacts on society. Information management and its sharing requires collaboration among planning and responding agencies such as FEMA. However, the lack of mechanisms to collect and share consistent data between federal, state, local, and other private entities present a barrier to effective collaboration. This can be attributed to the use of different systems for managing large amounts of data and information, thus complicating efforts towards standardized information management and knowledge-sharing. Ontologies present a domain of knowledge, which defines domain concepts and relationships between the concepts (Ontotext, 2024). Ontologies can be used to organize unstructured data, such as text data into a formal conceptualization of a particular domain (Antoniou and van Harmelen, 2008). However, the process of creating hazard ontologies is a time- and effort-intensive process. Ontology learning has been used to automate the construction of ontologies through the development of automated techniques to extract terms, synonyms, concepts, taxonomies, etc. from different data sources (Wohlgenannt and Minic, 2016). This research focuses on identifying and applying such methods to automate the development of hazard-specific ontologies from knowledge bases of disaster-related information (e.g., scholarly articles).

2 Ontology Learning System (OLS)

An Ontology Learning System (OLS) is a large and complex framework that encompasses various steps including data processing and information extraction (Wohlgenannt and Minic, 2016). The steps within the (OLS) are the following:

1. The initial step involves compiling a list of various sources of expert knowledge, including academic papers, technical reports, and authoritative web resources such as government websites.

2. Preprocessing techniques, as highlighted by Contreras et al. (2020), are applied to the collected documents in preparation for subsequent steps. **Text cleaning** removes unnecessary characters or symbols. **Word tokenization** splits the text into individual words or tokens. **Stop words**, which are commonly occurring words with little semantic significance, are removed from the text. Lastly, further refinement is carried out by **removing single and multiple characters** from the text

as required. This step helps eliminate noise or irrelevant information that could hinder the accuracy of the output ontology.

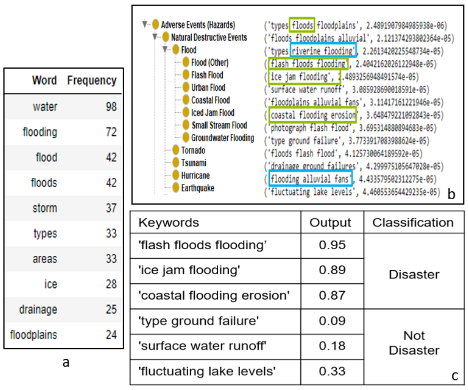
3. The next step involves extracting the most frequent words from the documents to identify commonly occurring terms. This is achieved using the FreqDist library (Chang et al., 2022), which analyzes the frequency distribution of each word in the document.

4. The most relevant words are then identified using Yet Another Keyword Extractor (YAKE), a keyword extraction algorithm that generates a list of the most relevant keywords in a body of text (Campos et al., 2020). This library does not require training on a specific set of documents and it can automatically prioritize a user-specified number of the most relevant terms using an unsupervised algorithm (Campos et al., 2020). This set of keywords represents a summary of the document. This is an important step toward developing an ontology for a specific domain. Campos et al. (2018) describe components used by YAKE, which include text preprocessing, feature extraction (emphasis through casing, importance through word position and frequency, context and structure through word relatedness and word difSentence), individual term scoring, and generating 3-keyword combinations (3-grams) with lower scores indicating more meaningful keywords. Candidate keywords are selected based on these scores, followed by data deduplication to avoid redundancy. The resulting set of keywords forms the final selection for use in subsequent steps.

5. Lastly, the keywords need to be classified to determine which terms should be included in the ontology. This decision-making process ensures that the resulting ontology contains the most appropriate and meaningful terms based on the extracted keywords. To facilitate this decision-making process, the keywords first need to be identified as being disasters or not using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). Values above 0.5 are considered "Disasters," and those below are labeled as "Not a Disaster." A multiclass classification approach then categorizes disaster-related keywords into specific types (e.g., flood or hurricane). Note, the training dataset used for BERT and the multiclass classification approach was obtained from a publicly available classified dataset on disaster tweets (Wiegmann et al., 2020). It includes 327,436 tweets across ten classes: other, hurricane, earthquake, flood, tornado, societal, industrial, wildfire, biological, and transportation.

3 Results

The FEMA technical report results (Figure 1-a) display a frequency table with "water," "flooding," and "flood" as the top three terms. In Figure 1-b, the green squares highlight keywords aligning with ontology categories from manual flood ontology created based on the literature. Blue squares suggest keywords for potential additional categories not in the ontology. Figure 1-c outcomes classify terms like "Flash Floods Flooding" and "Ice Jam Flooding" as "Disaster," and "Type Ground Failure" as "Not a disaster," achieving 89% accuracy, precision, and recall in disaster classification. The use of a multiclass classification approach was evaluated for its suitability in classifying disaster-related keywords as specific types of disasters (e.g., flood or hurricane). Preliminary results using a training dataset and three algorithms (Naive Bayes, Logistic Regression, and Linear Support Vector Classifier (SVC)) yielded the following test accuracy results for each of the ten classes listed above: Naive Bayes at 90%, Logistic Regression at 97%, and Linear SVC achieving the highest accuracy of 98.4%.



Word	Frequency
water	98
flooding	72
flood	42
floods	42
storm	37
types	33
areas	33
ice	28
drainage	25
floodplains	24

```
('types floods floodplains', 2.4891907984985938e-06)
('floods floodplains alluvial', 2.121374293802364e-05)
('types riverine flooding', 2.2613420225548734e-05)
('flash floods flooding', 2.4042162026122948e-05)
('ice jam flooding', 1.4893256948491574e-05)
('surface water runoff', 3.085928690018591e-05)
('floodplains alluvial fans', 3.11417161221946e-05)
('coastal flooding erosion', 3.648479221092843e-05)
('photograph flash flood', 3.695314880894683e-05)
('type ground failure', 3.7733917083988624e-05)
('floods flash flood', 4.125730064189592e-05)
('drainage ground failures', 4.299975105647028e-05)
('flooding alluvial fans', 4.433579502312275e-05)
('fluctuating lake levels', 4.460553654429235e-05)
```

Keywords	Output	Classification
'flash floods flooding'	0.95	Disaster
'ice jam flooding'	0.89	
'coastal flooding erosion'	0.87	
'type ground failure'	0.09	Not Disaster
'surface water runoff'	0.18	
'fluctuating lake levels'	0.33	

Figure 1: Results for FEMA Technical Report

Limitations

The results of the system are based on either one technical report or a compilation of six articles. We plan to expand our study to include a larger dataset in the future. It's important to note that ontologies in general are developed by independent parties, leading to variations in their structures (Halevy, 2005). This makes it challenging to validate results using a common standard ontology. Additionally, our findings haven't been validated by an expert in the field. Lastly, our system can identify keywords and concepts for building ontologies but lacks the ability to understand the semantic relationships between them. These limitations highlight the need for improvements in our approach and system capabilities for more reliable results.

References

Grigoris Antoniou and Frank van Harmelen. 2008. *A Semantic Web Primer, 2nd Edition (Cooperative Information Systems)*. The MIT Press.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Celia Nunes, and Adam Jatowt. 2018. A text feature based automatic keyword extraction method for single documents. In *European Conference on Information Retrieval*.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

CDC. Precipitation extremes: Heavy rainfall, flooding, and droughts | CDC [online]. 2020.

Siwei Chang, Ming-Fung Francis Siu, Heng Li, and Xiaowei Luo. 2022. Evolution pathways of robotic technologies and applications in construction. *Advanced Engineering Informatics*, 51:101529.

Jennifer O Contreras, Melvin A Ballera, and Enrique D Festijo. 2020. Ontology learning using hybrid machine learning algorithms for disaster risk management. In *Proceedings of the 2020 3rd International Conference on Signal Processing and Machine Learning*, pages 13–20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alon Halevy. 2005. Why your data won't mix: New tools and techniques can help ease the pain of reconciling schemas. *Queue*, 3(8):50–58.

Ontotext. What are ontologies? [online]. 2024.

Matti Wiegmann, Jens Kersten, Friederike Klan, Martin Potthast, and Benno Stein. Analysis of filtering models for disaster-related tweets [online]. 2020.

Gerhard Wohlgenannt and Filip Minic. 2016. Using word2vec to build a simple ontology learning system. In *International Workshop on the Semantic Web*.