V.S. Anoop* and S. Asharaf

Extracting Conceptual Relationships and Inducing Concept Lattices from Unstructured Text

<https://doi.org/10.1515/jisys-2017-0225>

Received May 16, 2017; previously published online September 26, 2017.

Abstract: Concept and relationship extraction from unstructured text data plays a key role in meaning aware computing paradigms, which make computers intelligent by helping them learn, interpret, and synthesis information. These concepts and relationships leverage knowledge in the form of ontological structures, which is the backbone of semantic web. This paper proposes a framework that extracts concepts and relationships from unstructured text data and then learns lattices that connect concepts and relationships. The proposed framework uses an off-the-shelf tool for identifying common concepts from a plain text corpus and then implements machine learning algorithms for classifying common relations that connect those concepts. Formal concept analysis is then used for generating concept lattices, which is a proven and principled method of creating formal ontologies that aid machines to learn things. A rigorous and structured experimental evaluation of the proposed method on real-world datasets has been conducted. The results show that the newly proposed framework outperforms state-of-the-art approaches in concept extraction and lattice generation.

Keywords: Formal concept analysis, concept extraction, concept lattices, relation extraction, knowledge discovery.

1 Introduction

Text is considered to be one form of data that is very rapidly generating because of the number of text-producing and text-consuming applications. User applications and platforms such as online social networks, digital libraries, e-commerce websites, and blogs generate text data, and this caused the creation of large unstructured text archives in organizations. These repositories are gold mines for organizations, as they contain invaluable patterns that help them leverage knowledge that can be used as an input to strategic but intelligent decision-making process. As the complexity and quantity of text data being generated grows exponentially, the need for more intelligent, scalable, and text-understanding algorithms is indispensable. The advent of semantic web, an extension and a meaning aware version of current World Wide Web, leads way to the introduction of numerous tools and techniques for leveraging, organizing, and presenting knowledge. Ontologies are building blocks of any meaning aware or semantic computing paradigm that comprises a set of concepts and their hierarchies and relationships in a domain of interest. Thus, the automated concept hierarchy learning from unstructured text has gained significant attention among text mining and natural language processing (NLP) researchers and practitioners. Concept hierarchy learning algorithms extract concepts from text and connect those concepts using potential relations that exist among them. Such hierarchies may find useful applications in concept-based ontology generation [19], concept-guided document summarization [24], and concept-guided information retrieval [10] to name a few.

*Corresponding author: V.S. Anoop, Data Engineering Lab, Indian Institute of Information Technology and Management-Kerala (IIITM-K), Thiruvananthapuram, India, e-mail: anoop.res15@iiitm.ac.in

S. Asharaf: Indian Institute of Information Technology and Management-Kerala (IIITM-K), Thiruvananthapuram, India

1.1 Contributions

This work proposes a framework for identifying commonly occurring relations that connect concepts in unstructured text data and then learn them using machine learning techniques. Specifically, the proposed approach can identify and learn subsumption (“is-a”), Hearst patterns [14] (“such as”, “or other”, “and other”, “including”, “especially”, etc.), and other potential indications of relations among concepts. This approach makes use of formal concept analysis (FCA) [34], which is a well-established mathematical theory of analyzing data, to form context table and concept lattices [34] from the identified concept and relations. These lattices can then be used for generating ontologies that may be used by intelligent and meaning aware computing systems. The authors compared the performance of this proposed system to some state-of-the-art methods through rigorous experiment, and results indicate that this approach outperforms the chosen baselines.

1.2 Organization

The rest of this paper is organized as follows. Section 2 discusses some of the very recent state-of-the-art methods in relation extraction (RE) and knowledge representation using FCA. The research objective and a formal problem definition are given in Sections 3 and 4, respectively. The new method proposed in this paper is explained in Section 5, and our experimental setup is given in Section 6. A detailed evaluation of results is given in Section 7, and we draw conclusions and then discuss future work in Section 8.

2 Related Work

RE is a subtask information extraction (IE) that aims at extracting relevant and potentially useful patterns or information from humongous data that are being generated day by day. The sheer volume and heterogeneity of data makes it difficult to analyze and extract these patterns manually. Thus, we need automated techniques for this process. NLP is one of the major areas that address this issue by scanning natural language texts and extract useful patterns. IE tasks, specifically the RE, have a long history and go back to late 1970s. However, successful commercial systems were introduced in the 1990s. In this section, we discuss some of the recent approaches in RE. In addition, we also throw light on state-of-the-art methods in FCA-based concept lattice generation approaches.

Informally, we can group all the RE approaches introduced in the literature into five categories: hand-built patterns, bootstrapping methods, supervised methods, distant supervision, and unsupervised approaches. Hand-built patterns uses handcrafted rules for extracting potentially relevant relation words from text, and one very notable work was introduced by M.A. Hearst known as Hearst patterns [14]. One issue with this approach is that it is difficult to write all sets of possible rules, and for other tasks such as meronym extraction, the set of rules will be different. But still, a good number of extensions are being reported that use Hearst patterns as their foundation for RE [16, 27, 29].

Another category of RE is the bootstrapping-based approach in which a specific set of seed relation instances are created and these are used for searching for new tuples. One such approach for extracting “author book” relation was the DIPRE [8]. Later, another system was introduced that uses the idea of bootstrapping; Snowball [1] extracts “organization-location” relation pairs. The limitations with the above-said algorithms are the specific relations they can only deal with. Users have to specify the type of relation they need to work with such as “author-book” or “organization-location”.

Later, TextRunner [36] was introduced in the domain of RE, which can learn relations, classes, and entities from a corpus in a self-supervised manner. This approach first tags the training data as positive and negative and then trains a classifier on the data to generate potential relations and entities. Another two-stage

bootstrapping algorithm [33] was proposed by Sun. In the first step, the algorithm uses a bootstrapping method to scan the tuples, and in the second stage, it learns relation nominals and contexts.

More recently, supervised and semisupervised approaches are found to be promising, and a very good number of literatures have been reported in the area of RE that uses deep learning techniques for identifying relation patterns. Very recently, a neural temporal RE approach [12] has been introduced in which the authors experimented with neural architectures for temporal RE. They showed that neural models that take tokens only outperform state-of-the-art hand-engineered feature-based models in performance. They also reported that encoding relation arguments with XML tags performs superior than a traditional position-based encoding. Another notable approach was introduced that attempts neural RE with selective attention over instances [21]. This work employs convolutional neural networks (CNN) to embed the semantics of sentences. Experimental results show that this model could make full use of all informative sentences and achieved significant and consistent improvement on RE task.

An approach for extracting relationships from clinical text was introduced [28] that exploited CNN to learn features automatically that reduces the dependency on manual feature engineering. They showed that CNN can be effectively used for RE in clinical text without being dependent on expert knowledge on feature engineering. Our proposed RE method uses machine learning techniques to classify Hearst patterns [14] (“such as”, “or other”, “and other”, “including”, “especially”, etc.) and other potential indications of relations such as “is-a” among textual concepts.

In recent years, FCA [34] has attained significant interest from research communities of various domains. FCA can analyze data that describe relationships that exist between a particular set of objects and their attributes. FCA is being widely used as a knowledge representation framework, especially in knowledge engineering and ontology generation tasks in information science. This proposed work also uses FCA to create concept lattices that incorporate a set of concepts and the relationships that connect them. Here, we discuss some of the very recent works on FCA, which use concept lattices for knowledge representation and ontology generation.

One of the recent notable works on extending FCA for association rule mining for knowledge representation is FCA-ARMM [15]. The authors integrated FCA and association rule mining model (ARMM) and developed a tool called FCA Miner, which is capable of generating association rules from real datasets. A portal retrieval engine based on FCA (PREFCA) [23] was introduced in which a portal’s semantic data were collected and formed in a concept lattice. Later in the information retrieval phase, a ranking has been done to retrieve the best result. Another work on identifying and validating ontology mappings using FCA was reported very recently [37]. The authors proposed a method called FCA-Map, which constructs formal contexts and then extracts mappings from the lattices derived. Then, a relation-based formal context is built and used for discovering additional structure mappings. An interactive knowledge discovery and data mining on genomic data using FCA [13] was introduced recently. They used FCA-based biclustering methods to index external databases for observing the evolution of genes throughout the different biclusters.

Very recently, an approach was proposed by Monnin et al. [22] which builds an optimal lattice-based structure for classifying RDF resources with respect to their predicates. The authors introduced the notion of lattice annotation that enabled to compare their classification to an ontology schema for confirming axioms that exhibit the subsumption relation or for suggesting a completely new one. The authors used the DBpedia dataset for their experiments and the results showed that their proposed approach could strongly demonstrate the ability of FCA to guide a possible structuring of Linked Open Data [22].

An approach for concept lattice reduction using fuzzy k-means clustering was introduced by Kumar and Srinivas [17]. The authors took into consideration the complexity in computing all the concepts from a large incident matrix and used a fuzzy k-means clustering for reducing the size of concept lattices. They also showcased the usefulness of their proposed method on two real-world applications such as information retrieval and information visualization. This method performed well on large context tables and the authors could represent reduced concept lattices efficiently [17]. A fuzzy clustering-based FCA for association rules mining was proposed by Kumar [18]. The author has done an association rule mining on a reduced formal context using fuzzy k-means clustering approach introduced in their previous work [17]. The authors conducted

experiments on two real-world healthcare datasets and showed that better association rule mining is possible on a reduced concept lattice [18].

Zhao and Zhang [37] introduced a novel method for identifying and validating ontology mappings using FCA in which the authors constructed three types of formal contexts and extracted mappings from the lattices derived. First, they showed that class names, labels, and synonyms share lexical tokens that may lead to lexical mappings across ontologies. Then, they proposed how this lattice can lead to validate the lexical matching to either positive or negative based on the lexical anchors. In the third phase, they showed how we can discover additional structural mappings from the positive relation-based context [37]. The authors conducted experiments and evaluated their methods on anatomy and large biomedical ontologies track of OAEI 2015 [37].

A very comprehensive survey on FCA and its research trends and applications was reported in the literature, which was compiled by Singh et al. [32]. The work is a torchbearer for researchers who wish to work on FCA and related areas. The authors summarized more than 350 recent research papers published after 2011 and indexed in major reputed indexing services. They specifically provided the mathematical foundations of each extension of FCA such as FCA with granular computing, intervalued FCA, and possibility theory [32]. Semenova and Smirnov [31] recently published a paper on building formal ontologies from incomplete data. They presented new models and methods for ontological data analysis, which facilitates the identification of conceptual structures or formal ontologies of a particular knowledge domain. They proposed an intelligent analysis of the incomplete data for building conceptual structures using FCA [31].

In this work, we make use of FCA for building context tables depicting various concepts and their associated relations and then transform this context into concept lattice. We show that efficient knowledge representation is thus possible and this may be extended into ontology engineering task.

2.1 Background: FCA

FCA is defined as a mathematical model or framework based on lattice theory [35], which is well suited for knowledge engineering and processing tasks. In recent years, the complexity and amount of data being produced across organizations has grown exponentially, and practitioners and researchers use FCA as an intelligent data analysis tool. In a basic setting, FCA generates two outputs for any given context table. The first one is called a concept lattice and the second one is called attribute implications. The former, the concept lattice, is a partially ordered collection of objects and its attributes and the latter, the attribute implications, describes particular attribute dependencies that are true in the context table [6]. One useful feature of FCA worth mentioning is that we can perform reasoning with dependencies in data, reasoning with concepts in data, and visualization of data with concepts and relationships. Some common examples are hierarchical arrangements of web search results, gene expression data analysis, analysis of organization of annotated taxonomies, etc. [6].

Definition 1: Formal context: In FCA, a formal context can be defined as a triplet $\langle X, Y, R \rangle$, where X and Y are nonempty sets and R is a binary relation between X and Y . For a formal context, elements x from X are called objects and elements y from Y are called attributes.

Definition 2: Concept forming operators: For a formal concept, $\langle X, Y, R \rangle$, operators $\uparrow: 2^X \rightarrow 2^Y$ and $\downarrow: 2^Y \rightarrow 2^X$ are defined for every $A \subseteq X$ and $B \subseteq Y$ by

$$A^\uparrow = \{y \in Y \mid \text{for each } x \in A: \langle x, y \rangle \in R\} \text{ and}$$

$$B^\downarrow = \{x \in X \mid \text{for each } y \in B: \langle x, y \rangle \in R\}.$$

Definition 3: Formal concept: In FCA, a formal concept in $\langle X, Y, R \rangle$ is a pair $\langle A, B \rangle$ of $A \subseteq X$ and $B \subseteq Y$ such that $A^\uparrow = B$ and $B^\downarrow = A$. For a formal concept $\langle A, B \rangle$ in $\langle X, Y, R \rangle$, A and B are called the extent and intent of $\langle A, B \rangle$, respectively.

Definition 4: Attribute implications: In FCA, an attribute implication can be defined as an expression $A \rightarrow B$, where $A, B \subseteq Y$, and it holds in a formal context if $A^\downarrow \subseteq B^\downarrow$. It means that any object that has all the attributes in A has also all the attributes in B . It is also well known that the sets of attribute implications that are satisfied by a context satisfy Armstrong's axioms [4].

FCA makes use of formal context (Definition 1) for data analysis in which each row corresponds to attributes and the filled value denotes the relationship between them. FCA takes this formal context as input and then outputs concept lattice that reflects generalization and specialization between the derived formal concepts from the incidence matrix [11]. These formal concepts are extensively used for knowledge processing tasks containing distinct extents and intents (set of objects and their relationships). These relations are represented in the form of a formal context, $F = (X, Y, R)$, where X is a set of objects, Y is a set of attributes, and R is a binary relation between them. From this given context, FCA creates a set of objects (A) and the set of all attributes (B) that are common for these objects.

Concept lattice: The concept lattice that is built from incidence matrix (context table) determines the hierarchy of formal concepts that follow a partial ordering principle as $(A_1, B_1) \leq (A_2, B_2)$ iff $A_1 \leq A_2$ ($B_2 \leq B_1$) and then give generalization and specialization between the concepts. That is, (A_1, B_1) is more specific than (A_2, B_2) . The attribute implications are represented in the form of $A \rightarrow B$ over the set Y . There are several algorithms developed for generating concept lattices [5, 9, 20, 25]. An example of a formal context showing airlines and their sector of operations and corresponding concept lattice visualization are shown in Figures 1 and 2, respectively. In the formal context (Figure 1), the rows represent the concepts or objects (in this case, "Air Canada", "Air New Zealand", and "Air India") and the columns represent the set of attributes (in this case, "Latin America", "Asia", "Europe", and "Middle East"). A cross ("X") in the intersection (cell) on the formal context denotes that the object has a corresponding property. In Figures 1 and 2, it denotes that an airline is operating in that particular sector. See Ref. [6] for a more detailed and comprehensive explanation of FCA and its related theory.

A	B	C	D	E
	Latin Amer...	Asia	Europe	Middle East
Air Canada	X		X	
Air New Zealand	X	X		
Air India		X		X

Figure 1: Formal Context Showing Airlines and their Sector of Operations.

Figure 2: Concept Lattice Generated for Formal Context Given in Figure 1.

3 Research Objective

The following are our main research objectives:

1. Introduce the task of RE from unstructured text and major approaches and categories for the same.
2. Propose a framework that uses machine learning approach for automatically extracting and learning subsumption relation (“is-a”), Hearst patterns (“such as”, “or other”, “and other”, “including”, “especially”, etc.), and other potential indications of relations among concepts.
3. Represent the knowledge (concepts and extracted relationships) using FCA.
4. Verify experimentally the effectiveness of the method in extracting and representing real-world concepts and relationships.

4 Problem Definition

We now define the problem formally. RE problem is the task of detecting and classifying semantic relationships that connect entities or phrases or concepts in a corpus of interest. Given a static document corpus D , the relationship extraction task identifies valid relation words that connect two concepts together. For example, consider the sentence, “Alzheimer is a degenerative disease”. The words “Alzheimer” and “degenerative disease” are potential concepts in a medical text document. The relationship extraction method identifies “is-a” as a potential relation that connects these two concepts. Given a static document corpus $D = d_1, d_2, \dots, d_n$ that contains key-phrases or concepts $C = c_1, c_2, \dots, c_n$, our problem is to identify semantically distinguishable relations that connect $c_1, c_2, \dots, c_n$. We also address the problem of representing this knowledge using FCA, which is a widely used knowledge representation framework that comes with well-implemented mathematical models.

5 Proposed Approach

In this section, we outline our proposed approach for identifying and extracting relationships that connect entities or concepts that are extracted from unstructured text. Considering the complex sentence and language structures, concept extraction as well as RE is an extremely difficult task in NLP and information retrieval. Although there are many attempts reported on how to extract them, majority of the works are heavily dependent on the specific corpus chosen for the experiment. In previous works, we have also attempted the concept extraction task, which is guided by a topic modeling process that suits well on any plain text corpus [2, 3]. In this work, our main focus is on the RE task; thus, the process of concept extraction is not emphasized. Here, we are using an off-the-shelf tool for identifying potential entities, phrases, and concepts from our static document corpus and then implement our relationship extraction algorithm on top of it to extract relation patterns. An overall workflow of the proposed approach is shown in Figure 3.

Figure 3: Overall Workflow of the Proposed Approach.

6 Experimental Setup

In this section, we describe our experimental setup and a detailed evaluation of the results. Two separate experiments are conducted. The first one is for extracting semantically valid relationships that connect concepts extracted. The second one is for validating the usefulness of FCA for representing concepts and relationships extracted from a plain text corpus. The entire dataset, noun phrases, and the Python code are available in an open repository and can be freely accessed at <https://github.com/anoop-research/relation-extraction>.

6.1 Dataset Description

For the experiment, we have used disease description data that are publically available in unstructured form from the Medscape website (<http://www.medscape.com>), which is an online global destination for physicians and other healthcare professionals. This website offers the latest medical news, expert opinions, and disease details. We have crawled the website for diseases and treatment descriptions for two categories (cardiology and neurology) and collected data stored in plain text files. Some of the concepts or medical phrases extracted from those plain text files are shown in Table 1. For the cardiology category, we have collected descriptions of 45 diseases, such as acute coronary syndrome, alcoholic cardiomyopathy, heart failure, and hypertension, and for neurology, there are 42 descriptions for diseases, such as Parkinson’s disease, depression, and Alzheimer’s disease. A snapshot of such a description (for Parkinson’s disease) is shown in Figure 4.

Table 1: Some of the Concepts/Medical Phrases Extracted for Cardiology and Neurology Categories.

Cardiology	Neurology
Acute aortic dissection	Central nervous system
Marfan syndrome	Potential toxic metabolites
Type III dissections	Pathological processes
β-Adrenergic blocker	Progressive disability
Thoracic aortic dissections	Vascular parkinsonism
Atherosclerotic disease	Thalamocortical pathway
Lymphocyte activation	Painful muscular contractions
Urine microalbumin	Umbilical cord contamination
Rheumatogenic strains	Neonatal tetanus
Streptococcal infections	Elastic membrane

Parkinson disease (PD) is one of the most common neurologic disorders, affecting approximately 1% of individuals older than 60 years and causing progressive disability that can be slowed, but not halted, by treatment. The 2 major neuropathologic findings in Parkinson disease are loss of pigmented dopaminergic neurons of the substantia nigra pars compacta and the presence of Lewy bodies and Lewy neurites. See the images below. Initial clinical symptoms of Parkinson disease include the following: Onset of motor signs includes the following: Nonmotor symptoms Nonmotor symptoms are common in early Parkinson disease. Recognition of the combination of nonmotor and motor symptoms can promote early diagnosis and thus early intervention, which often results in a better quality of life. See Clinical Presentation for more detail. Parkinson disease is a clinical diagnosis. No laboratory biomarkers exist for the condition, and findings on routine magnetic resonance imaging and computed tomography scans are unremarkable.....

Figure 4: Snapshot of Disease Description Collected from Medscape (<http://www.medscape.com>).

6.2 Dataset Preprocessing

Dataset preprocessing concentrates on tidying up the data by removing unwanted characters, words, special symbols, and links from external sources. We have not removed stop-words from the corpus for this experiment as some words in the stop-words list may be found useful in identifying a particular relation word, for example, “is-a”. Special symbols and other irrelevant characters are removed using regular expressions, and we used Snowball stemmer for reducing the words to its root word, say, “affecting” to “affect”. Then, we have vectorized these words to feed into our proposed machine learning model.

6.3 Building a Machine Learning Model

Our next step is to create a machine learning classifier that learns the Hearst patterns and other potential indications of relation patterns. We have used many classification algorithms to build the model, such as decision tree, random forest, Adaboost, and support vector machine (SVM). The entire dataset has been split into 70–30, where 70% is for training and 30% for testing the model. For training and testing, we have used a server machine configured with AMD Opteron 6376 with 2.3 GHz speed and 16 core processor and 16 GB of main memory. For implementing the classifier, Python 2.7 version is used along with “sci-kit learn” [26] library. The input to these classifiers is in the form of a sentence-relation word matrix where the concept/keyphrase tagged sentences are in rows of the matrix and the relation words, such as “is-a” and “such as”, are along the column. The cell contains the total count of a specific relation word that occurred in that sentence. As this is a binary classifier, “0” or “1” is given as the label based on the absence or presence of the desired relation. For this experiment, the input matrix contains 10,000 such rows, which are given as input to the four relation classifier models. After training, for testing, we present a new sentence to the trained model, and the model will output either “1” or “0” based on the presence or absence of the relation that will be converted to a formal context for building concept lattices.

For decision tree classifier, “gini” criterion is used as a measure of quality of split, and the maximum depth parameter has been chosen as 15 by a trial-and-error method. Second, random forest classifier has been implemented with the number of estimators as 300, maximum depth as 15, and random state as 42. Our Adaboost classifier used decision tree classifier as the base estimator that has a maximum depth of 8 and random state of 42. The learning rate parameter, the number of estimators, and the random state were chosen to 0.9, 500, and 1332, respectively. For our SVM classifier, we have set the number of random state as 22 and maximum number of iterations as 100.

6.4 Creating Concept Lattices

Once major concepts, phrases, and relationships have been extracted, the knowledge may be represented in an intuitive and informative way so that meaning aware applications can be built on top of the knowledge. We used FCA for deriving concept hierarchy or formal ontology that represents concepts and associated relationships we have leveraged. To build the lattice, a table with logical attributes represented as a triplet $\langle X, Y, R \rangle$ should be created first. In the triplet $\langle X, Y, R \rangle$, R denotes a binary relation between the objects X and Y . In our case, X denotes a set of disease names and Y denotes a set of attributes of the disease. For example, consider “hypertension” as the case; then, we have “blood pressure”, “breathing disorder”, “cortisol stress reactivity”, etc., as its attributes. The binary relation R has a value of 1 if “hypertension” has a particular attribute or else the value will be 0. The size of the entire formal context and the concept lattice is so big. For the “cardiology” domain, the formal context contained 45 objects (disease names) and 7746 attributes (symptoms). For the “neurology” domain, there are 42 objects (disease names) and 9156 attributes (symptoms). Due to space constraints, it is impossible to show the entire context table and concept lattice that has been generated for the entire dataset. A part of such a context table created is shown in Figure 5 and the corresponding concept lattice is shown in Figure 6.

A	B	C	D	E	F
	blood pres...	breathing ...	cytotoxic eff...	fulminant ...	genetic trait
hypertension	X	X			
myocarditis			X		
nephrosclerosis				X	
pulmonic stenosis					
brown syndrome					X

Figure 5: Part of a Formal Context Generated from the Original Dataset Using Our Proposed Method.

Figure 6: Part of a Concept Lattice Generated from the Original Dataset Using Our Proposed Method.

6.5 Baselines Chosen

We have chosen two baselines for comparing our proposed method. Of all classifier algorithms chosen (decision tree, random forest, Adaboost, and SVM), SVM showed better accuracy on relation classification. Thus, we compared the SVM model to the chosen baselines. We have chosen two baselines.

- Baseline 1 [7]: This approach extracts facts from natural language texts with conceptual modeling [7]. This work shows the application of FCA for extracting facts from natural language text. Their approach combines concept graphs and concept lattices for leveraging facts, which is closely associated with our proposed approach. They use concept lattices to model relationships that connect words and then these relationships for interpreting formal concepts as possible facts.
- Baseline 2 [30]: The second baseline, on the contrary, attempts to create a public dataset containing more than 400 million hypernymy relations from CommonCrawl web corpus. Although we are not considering all the relations used in their work, we have chosen this as our second baseline, as our proposed work is aligned to their workflow in a greater extent.

A comparison of the results of these baselines to our proposed framework and a detailed evaluation are discussed in Section 7.

7 Results and Evaluation

This section describes the results of our rigorous and systematic experiment on relation classification and knowledge representation using FCA. As explained in Section 6, for the relation classification, we have chosen four different algorithms, such as Adaboost, random forest, decision tree, and SVM. The precision, recall, and F1 score reported by our machine learning classifier are shown in Tables 2–4. Of the four different

Table 2: Precision, Recall, and F1 Score of Different Classifier Algorithms on Our Dataset.

Algorithm	Precision	Recall	F1 score
Adaboost	0.95	0.85	0.90
Random forest	0.97	0.86	0.91
Decision tree	0.95	0.84	0.89
SVM	0.98	0.87	0.92

Table 3: Precision, Recall, and F1 Score Comparison of Baselines and Our Proposed Method for RE.

Algorithm	Precision	Recall	F1 score
Baseline (for RE) [7]	0.79	0.81	0.79
Proposed	0.91	0.87	0.88

Table 4: Precision, Recall, and F1 Score Comparison of Baselines and Our Proposed Method for Concept Lattice Generation.

Algorithm	Precision	Recall	F1 score
Baseline (for lattice generation) [30]	0.80	0.78	0.78
Proposed	0.88	0.87	0.87

Figure 7: Graph Representation of Classifier Performance on the Chosen Dataset.

algorithms chosen with optimal parameters, SVM is found to show better classification accuracy; thus, this model has been chosen for comparing the performance of our proposed method to the chosen baselines. The normalized confusion matrix for all the four classification algorithms is shown in Figure 7 and the classification accuracy comparison in terms of precision, recall, and F1 score in a graph is shown in Figure 8.

8 Conclusions and Future Work

This paper proposed a framework for extracting relationships that connect concepts and phrases found in unstructured text documents. We make use of a machine learning-based approach for learning commonly occurring relations such as “is-a” and other patterns discussed in Hearst patterns such as “such as”, “or other”, “and other”, “including”, and “especially”. This approach employs different machine learning algorithms to classify potential relationships, such as Adaboost, random forest, decision

Figure 8: Normalized Confusion Matrix for (A) Adaboost, Random Forest (B), Decision Tree (C), and SVM (D) Classifiers.

tree, and SVM. This work makes use of FCA to represent noun-phrases and relations that are leveraged using our proposed RE algorithm. Experiments on real-world medical dataset collected from public web shows that this proposed method extracts better conceptual structures when compared to the baselines [7, 30].

As the end results are promising, our future work will be mainly on the direction of improving the accuracy of our classification engine and extracting more semantically valid relation patterns. This may generate more fine-grained facts from unstructured text and may aid ontology enrichment process in semantic computing paradigms. This current experimental setup will only work with static unstructured text corpus for extracting facts and building concept lattices. In the future, we may extend this for dynamically generated text contents from platforms such as social networks.

Acknowledgments: The authors thank all researchers from the Data Engineering Lab at the Indian Institute of Information Technology and Management-Kerala (IIITM-K) for their suggestions that improved the quality of this paper. The authors also acknowledge the anonymous reviewers for their constructive comments.

Bibliography

- [1] E. Agichtein and L. Gravano, Snowball: extracting relations from large plain-text collections, in: *Proceedings of the 5th ACM Conference on Digital Libraries*, pp. 85–94, San Antonio, TX, USA: ACM, 2000.
- [2] V. S. Anoop, S. Asharaf and P. Deepak, Learning concept hierarchies through probabilistic topic modeling, *Int. J. Inf. Process.* **10** (2016), 1–11.
- [3] V. S. Anoop, S. Asharaf and P. Deepak, Unsupervised concept hierarchy learning: a topic modeling guided approach, *Proc. Comput. Sci.* **89** (2016), 386–394.
- [4] W. W. Armstrong, Dependency structures of data base relationships, in: *IFIP Congress*, vol. 74, pp. 580–583, 1974.
- [5] E. Bartl, H. Rezankova and L. Sobisek, Comparison of classical dimensionality reduction methods with novel approach based on formal concept analysis, in: *International Conference on Rough Sets and Knowledge Technology*, pp. 26–35, Springer, Berlin/Heidelberg, 2011.
- [6] R. Belohlavek, *Introduction to Formal Concept Analysis*, Department of Computer Science, Palacky University, Olomouc, 2008.
- [7] M. Bogatyrev, Fact extraction from natural language texts with conceptual modeling, in: *International Conference on Data Analytics and Management in Data Intensive Domains*, pp. 89–102, Moscow, Russia: Springer, 2016.
- [8] S. Brin, Extracting patterns and relations from the world wide web, in: *International Workshop on the World Wide Web and Databases*, pp. 172–183, Springer, Berlin/Heidelberg, 1998.
- [9] V. Codocedo, C. Taramasco and H. Astudillo, Cheating to achieve formal concept analysis over a large formal context, in: *The 8th International Conference on Concept Lattices and Their Applications-CLA 2011*, pp. 349–362, LORIA Nancy, France, 2011.
- [10] C. Cui, J. Shen, Z. Chen, S. Wang and J. Ma, Learning to rank images for complex queries in concept-based search, *Neurocomputing*, Elsevier (2017, In Press).
- [11] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, Cambridge University Press, Cambridge, UK, 2002.
- [12] D. Dligach, T. Miller, C. Lin, S. Bethard and G. Savova, Neural temporal relation extraction, *European Chapter of the Association for Computational Linguistics*, p. 746, Valencia, Spain, 2017.
- [13] J. M. Gonzalez-Calabozo, F. J. Valverde-Albacete and C. Pelaez-Moreno, Interactive knowledge discovery and data mining on genomic expression data with numeric formal concept analysis, *BMC Bioinform.* **17** (2016), 374.
- [14] M. A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: *Proceedings of the 14th Conference on Computational Linguistics*, vol. 2, pp. 539–545, Association for Computational Linguistics, Nantes, France, 1992.
- [15] T. Herawan, M. M. Deris and A. R. Hamdan, FCA-ARMM: a model for mining association rules from formal concept analysis, in: *Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016)*, Bandung, Indonesia, August 18–20, 2016 *Proceedings*, vol. 549, p. 213, Springer, 2017.
- [16] T. Kawaumra, M. Sekine and K. Matsumura, Hyponym/hypernym detection in science and technology thesauri from bibliographic datasets, in: *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*, pp. 180–187, San Diego, CA, USA: IEEE, 2017.
- [17] C. A. Kumar and S. Srinivas, Concept lattice reduction using fuzzy k-means clustering, *Expert Syst. Appl.* **37** (2010), 2696–2704.
- [18] C. A. Kumar, Fuzzy clustering-based formal concept analysis for association rules mining, *Appl. Artif. Intell.* **26** (2012), 274–301.
- [19] N. Kumar, M. Kumar and M. Singh, Automated ontology generation from a plain text using statistical and NLP techniques, *Int. J. Syst. Assur. Eng. Manage.* **7** (2016), 282–293.
- [20] S. O. Kuznetsov and S. A. Obiedkov, Comparing performance of algorithms for generating concept lattices, *J. Exp. Theor. Artif. Intell.* **14** (2002), 189–216.
- [21] Y. Lin, S. Shen, Z. Liu, H. Luan and M. Sun, Neural relation extraction with selective attention over instances, in: *Proceedings of ACL*, vol. 1, pp. 2124–2133, 2016.
- [22] P. Monnin, M. Lezoche, A. Napoli and A. Coulet, Using formal concept analysis for checking the structure of an ontology in LOD: the example of DBpedia, in: *23rd International Symposium on Methodologies for Intelligent Systems, ISMIS*, 2017.
- [23] E. Negm, S. AbdelRahman and R. Bahgat, PREFCA: a portal retrieval engine based on formal concept analysis, *Inf. Process. Manage.* **53** (2017), 203–222.
- [24] H. Oliveira, R. Lima, R. D. Lins, F. Freitas, M. Riss and S. J. Simske, A concept-based integer linear programming approach for single-document summarization, in: *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pp. 403–408, Recife, Pernambuco, Brazil: IEEE, 2016.
- [25] J. Outrata and V. Vychodil, Fast algorithm for computing fixpoints of Galois connections induced by object-attribute relational data, *Inf. Sci.* **185** (2012), 114–127.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel and J. Vanderplas, Scikit-learn: machine learning in Python, *J. Mach. Learn. Res.* **12** (2011), 2825–2830.
- [27] S. Roller and K. Erk, Relations such as hypernymy: identifying and exploiting Hearst patterns in distributional vectors for lexical entailment, arXiv preprint arXiv:1605.05433 (2016).
- [28] S. K. Sahu, A. Anand, K. Oruganty and M. Gattu, Relation extraction from clinical texts using domain invariant convolutional neural network, arXiv preprint arXiv:1606.09370 (2016).

- [29] J. Seitner, C. Bizer, K. Eckert, S. Faralli, R. Meusel, H. Paulheim and S. Ponzetto, A large database of hypernymy relations extracted from the web, in: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, Portoroz, Slovenia, 2016.
- [30] J. Seitner, C. Bizer, K. Eckert, S. Faralli, R. Meusel, H. Paulheim and S. Ponzetto, A large database of hypernymy relations extracted from the web, in: *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*, Portoroz, Slovenia, 2016.
- [31] V. A. Semenova and S. V. Smirnov, Intelligent analysis of incomplete data for building formal ontologies, in: *CEUR Workshop Proceedings*, vol. 1638, pp. 796–805, 2016.
- [32] P. K. Singh, C. A. Kumar and A. Gani, A comprehensive survey on formal concept analysis, its research trends and applications, *Int. J. Appl. Math. Comput. Sci.* **26** (2016), 495–516.
- [33] A. Sun, A two-stage bootstrapping algorithm for relation extraction, in: *Proceedings of Recent Advances in Natural Language Processing*, pp. 76–82, Borovets, Bulgaria, 2009.
- [34] R. Wille, Restructuring lattice theory: an approach based on hierarchies of concepts, in: *Ordered Sets*, pp. 445–470, Springer, The Netherlands, 1982.
- [35] R. Wille, Concept lattices and conceptual knowledge systems, *Comput. Math. Appl.* **23** (1992), 493–515.
- [36] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead and S. Soderland, TextRunner: open information extraction on the web, in: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 25–26, Association for Computational Linguistics, Rochester, New York, 2007.
- [37] M. Zhao and S. Zhang, Identifying and validating ontology mappings by formal concept analysis, in: *Proceedings of the 15th International Semantic Web Conference*, pp. 61–72, Kobe, Japan, 2016.