A KNOWLEDGE GRAPH AND RULE BASED REASONING METHOD FOR EXTRACTING SAPPHIRE INFORMATION FROM TEXT

Bhattacharya, Kausik;
Chakrabarti, Amaresh

Indian Institute of Science Bangalore

ABSTRACT

Representation of design information using causal ontologies is very effective for creative ideation in product design. Hence researchers created databases with models of engineering and biological systems using causal ontologies. Manually building many models using technical documents requires significant effort by specialists. Researchers worked on the automatic extraction of design information leveraging the computational techniques of Machine Learning. But these methods are data intensive, have manual touch points and have not yet reported the end-to-end performance of the process. In this paper, we present the results of a new method inspired by the cognitive process followed by specialists. This method uses the Knowledge Graph with Rule based reasoning for information extraction for the SAPPhIRE causality model from natural language texts. Unlike the supervised learning methods, this new method does not require data intensive modelling. We report the performance of the end-to-end information extraction process, which is found to be a promising alternative.

Keywords: Computational design methods, Creativity, Ontologies, Information Extraction, Machine learning

Contact:
Bhattacharya, Kausik
Indian Institute of Science Bangalore
India
kausikb@iisc.ac.in

Cite this article: Bhattacharya, K., Chakrabarti, A. (2023) 'A Knowledge Graph and Rule Based Reasoning Method for Extracting Sapphire Information from Text', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.23

1 INTRODUCTION

1.1 Context

Stimuli are important in innovative product design (Howard et al., 2010). Causal representations of systems play a significant role in creative design ideation (Baldussu et al., 2012). Databases with causal ontology based models were developed to support the design process (Fu et al., 2014). However, compared to the massive volumes of design information stored as natural language documents of various types and structures, these databases have a limited number of models. It is humanly impossible to mine data from this vast number of unstructured documents in search of stimuli. Hence, design methods for automatically extracting design information are developed using machine learning techniques. However, these design methods are partially automated and have manual steps. Literature only reports the machine learning models' performance instead of the performance of the end-to-end process, including the manual steps. In this paper, we first reviewed the methods used in Information Extraction, including their applications in product design. We then looked into the application of the Information Extraction task to generate a causal representation of systems from natural language texts. From this review, we identified the opportunities for further research and framed our research question. We then presented the details of our work and results and ended the paper with a conclusion and future work.

1.2 Information extraction

Information extraction is "the process of acquiring knowledge by skimming a text and looking for occurrences of a particular class of objects and for relationships among objects" (Russel & Norvig, 2010). Cowie et al. (1996) describe the components of information extraction systems as (a) Filtering to determine the relevance of text based on word statistics, (b) Parts of Speech tagging of words, (c) Semantic tagging for recognizing major phrasal units, (d) Parsing to create semantic level maps showing relations between phrasal units (e) Discourse reference in merging inter-sentence level overlaps and finally (f) output generation using predefined output format. Three methodologies are primarily used in the information extraction tasks (Tang et al., 2008), namely, (a) Rule Learning based Extraction Methods using rules or a template to extract information - further subdivided into a dictionary-based method, rule-based method, and wrapper induction, (b) Classification based Extraction Methods treating Information Extraction as a classification problem and (c) Sequential Labelling based Extraction Methods - where a document is viewed as a sequence of tokens, and a sequence of labels is assigned to each token to indicate the property of the token. Hidden Markov Model, Maximum Entropy Markov Model and Conditional Random Field are widely used sequential labeling models. Work on knowledge discovery using text mining technique (Yang et al., 2018) defines a domain ontology and then, using that ontology carries out the knowledge extraction along with data visualization. Illustrative results are obtained using a corpus of 140,000 technical reports. The use of the Semantic web for information extraction is another method widely reported in the literature. Gangemi (2013) benchmarked 14 different tools against the standard information extraction tasks using the precision, recall and F scores of these tools for general tasks. The precision metric gives the % of positively predicted labels that are correct and the recall metric gives the % of all the positives that the model predicts correctly. F1-score is computed as the harmonic mean of recall and precision (Russel & Norvig, 2010).

Nédellec et al. (2009) argue that future effort in information extraction needs to formalize and reinforce the relationship between text extraction and the ontology model. This work explains a two-step process where the first step identifies the occurrences of the candidate semantic entities and then identifies the ontology specific relationship between them in the next step. It uses supervised Learning to train a model to determine the relationship between semantic units. Ontology-Based Information Extraction (OBIE) is a type of information extraction where the information extraction process uses ontologies. The output is presented in the form of an ontology, specified by a domain expert beforehand and is domain specific. Information extraction decides a suitable technique guided by ontology. The process often uses a semantic lexicon (e.g., WordNet) for the language in concern. Wimalasuriya et al. (2010) studied various ontology-based information extraction methods reported in the literature. Vargas-Vera et al. (2001) presented an ontology-based information extraction tool with an ontology-based mark-up component, allowing the user to mark up only relevant pieces of

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

information while browsing documents. Then there is a learning module that learns rules from these examples which are used later to identify and extract objects and the relationship between objects.

As presented above, several methods are developed and used for a variety of information extraction tasks, such as named entity recognition, terminology extraction, sense disambiguation, taxonomy induction, and relation extraction. However, the documents' heterogeneity and lack of standard structure make Information Extraction still an area of research in machine learning.

2 INFORMATION EXTRACTION FOR DESIGN REPRESENTATION

Designers often use stimuli in design. A detailed study to understand designers' approach in selecting inspirational stimuli in the conceptual design phase was conducted where the representation of external stimuli was an important factor in the inspiration process (Gonçalves et al., 2016). Since massive design information is captured in many design documents, technical reports and patents, researchers used information extraction techniques to extract useful information automatically and use it in design.

2.1 Knowledge graph based

Due to the advantage of a semantic web representing a vast amount of information in the form of a connected graph, it is used in many engineering applications (Han et al., 2022). Hsieh et al. (2011) developed a method for extracting concepts, instances and relationships from a handbook of the engineering domain. The extracted information is then converted into a semantic web. A co-occurrence graph is a powerful technique to represent semantic relations. Hence it is used to identify relevant technical parameters for a certain domain by comparing thesauri automatically extracted from patents (Cascini et al., 2011). A more generalized semantic network TechNet is specifically trained on larger technology related data sources compared to WordNet and ConceptNet, which are more broad based and lesser engineering centric (Sarica and Luo, 2021). A rule based technique is used to generate an engineering Knowledge Graph from Patent Database. This was found to have a greater size and coverage than the previously used knowledge graphs (Siddharth et al., 2022). In recent work (Zhang et al., 2020), a new method using a biologically inspired adaptive growth approach (BIAG) was developed to learn domain ontologies from engineering documents due to the similarity in domain ontology learning process and tree growth. BIAG is modeled using the genes and three stages of tree lifecycle: seed formation, root development and tree growth. All these methods give a powerful way of knowledge extraction but they do not follow any ontology based representation.

2.2 Ontology based

Researchers worked on developing ontology based information extraction for design applications using natural language processing and domain-specific design ontology. Li and Ramani (2009) presented a method to extract domain ontology using design themes (things that are of the designer's interest) from a set of technical documents. It uses syntax analysis, semantics analysis, and domain ontology to discover the concepts and relationships in those documents. This method produces higher precision and recall scores compared to the vector model, which uses the classical word vector method. Though this method uses domain specific ontology to extract information of interest, the extracted information is presented as a knowledge graph and not as a causal representation of the systems. Cheong & Shu (2012) developed a method to extract causally related functions for biological systems from natural language documents. This work uses a causal relation extraction algorithm to read the word's part-of-speech tags and dependency relations to identify the syntactic patterns. This approach reported higher accuracy compared to the baseline. The major limitation is that this captures causal relation only for functions and works at a single sentence level. Keshwani (2018) developed a semi-automated method using a support vector machine approach to extract words corresponding to the constructs of the SAPPhIRE model of causality given in Figure 1 (Chakrabarti et al., 2005, Venkataraman and Chakrabarti, 2009) from multi-sentence natural language descriptions of systems. It is a four-step process where all the relevant words are first identified in a pre-processing and the classifier then determines their SAPPhIRE label. The accuracy of this classifier is given in Table 1. Goel et al. (2020) presented another work on understanding design documents using the extracted information. It uses a machine learning approach to extract words belonging to the SBF, i.e., Structure-Behavior-Function (Goel et al., 2009) ontology. This work reported a performance comparison of multiple models. Table 2 summarizes the result reported in this work. The

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

last two research (Keshwani, 2018 and Goel et al., 2020) used a limited number of hand annotated data to train and validate models, and none reported a very high accuracy score for the classification. More importantly, these methods comprise multiple steps with manual touchpoints, but the accuracy of the end-to-end process was not reported.

Table 1. Evaluation metrics for the classifier with test data (Keshwani, 2018)

Evaluation Metrics	Linear SVM Classifier (at c = 2)
Precision	0.74
Recall	0.68
F-Measure	0.70
Stratified Cross Validation Accuracy for entire dataset (k = 10)	0.72 ± 0.07

Table 2. Comparison of models (Goel et al., 2009)

Classifier	Weighted Accuracy	Precision	Recall	F-Beta Score
SVM	72.8%	0.81	0.72	0.76
Decision Tree	70.9%	0.79	0.71	0.74
Gradient Tree	86.4%	0.83	0.86	0.84
Random Forest	76.5%	0.83	0.76	0.79
Logistic Regression	69.1%	0.82	0.69	0.74

3 RESEARCH QUESTIONS

Analogical reasoning is considered the core of human cognitive activities responsible for creativity in product design (Goel and Shu, 2015). Causal ontologies, such as Structure-Behavior-Function (Goel et al., 2009) and SAPPhIRE (Chakrabarti et al., 2005), are very effective for analogical reasoning. We saw that the current methods of extracting information for causal ontologies are based on the classical supervised learning approach. But these methods did not have adequate data for training and validation and their accuracy needs to be higher. Deep learning methods known for higher accuracy are not tried yet and require far more data for model building. Moreover, building the current models requires significant manual pre- and post-processing and the accuracy of the complete end-to-end process is not reported. So there is scope for more research to reduce, if not eliminate, dependency on a manual effort and improve the accuracy of the end-to-end process.

The research question therefore is:

- How to extract causal ontology from natural language systems description using a computational model that does not require any data or uses less data for model building yet gives the same, if not better, level of accuracy?

A new process using a Knowledge Graph and Rule based reasoning for extracting words from a natural language description, representing the constructs of the SAPPhIRE model of causality, was reported earlier (Bhattacharya et al., 2023). However, the previous work studied the process performance in reducing variability in the extracted information across multiple designers. In this paper, our goal is to study the accuracy of information extraction using Knowledge Graph and Rule based reasoning. Before presenting results, we explain the details of building a knowledge graph from text and the rules for reasoning to extract information, which were not shown in detail before.

4 DEVELOPING A METHOD TO EXTRACT CAUSAL ONTOLOGY

4.1 Approach

We adopted an approach that is close to the cognitive behavior of the specialists when they create the SAPPhIRE model from information given in a technical document (Bhattacharya et al., 2023). Specialists first read the document to understand the information entities given in the document and identify the causal relationship between those entities. This is essentially a Knowledge Graph connecting two entities by a relation. With the help of expert knowledge acquired from experience, a specialist then applies reasoning to categorize these information entities and their relationship per the definition of each SAPPhIRE construct. We created a set of rules that categorizes words in a knowledge graph into multiple categories. Each of these categories is mapped to a specific SAPPhIRE construct.

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

4.2 Creating knowledge graph

We use the Parts-of-Speech tag of the words to generate a knowledge graph representation of a sentence. Before processing, every compound sentence is broken down into independent sentences. Then in each sentence, we look for four types of knowledge graphs, namely, (a) Noun-*Verb*-Noun (e.g., Load - *completes* - circuit), (b) Noun-*Verb*-Adverb (e.g., 'Thermal Wheel' - *rotates* - slowly), (c) Noun-*Preposition*-Noun (e.g., circuit - *between* - terminals) and (d) Noun-*verb*-Adjective (e.g., Frame-*is*-Static (part)). It may be noted that sometimes, there may not be mention of any verb between a Noun and it's adjective. In such a situation, we assume a state of being verb between them. We then connect all the knowledge graphs using three conditions, namely (a) using the Proper Nouns, (b) using the sequence of Action Verbs (e.g., 'The anode *experiences* an oxidation reaction' and 'The reaction in the anode *creates* electrons') and (c) Conditions of the Action Verbs represented through an Adverb or a Conjunction. (e.g., "*When* a load completes the circuit *between* the two terminals, the battery produces electricity"). Appendix A shows an example of a Knowledge Graph for a paragraph with more than one sentence.

4.3 Rule based reasoning and mapping to the SAPPhIRE Model

We developed rules to categorize the knowledge capture in a knowledge graph. These rules are harvested based on how specialists use reasoning to understand natural language texts. For this Reasoning task, we use the POS tags of the words in the knowledge graph and apply the reasoning rules given in Table 3 to categorize the extracted information. Figure 2 shows how each information category in Table 3, represents a building block of the system's description based on the systems modeling knowledge (Chakrabarti et al., 2005; Hubka and Eder, 2012).

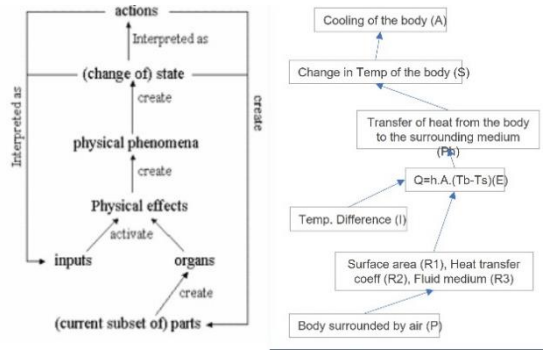


Figure 1. SAPPhIRE Model of causality with an example of heat transfer (Chakrabarti et al., 2005)

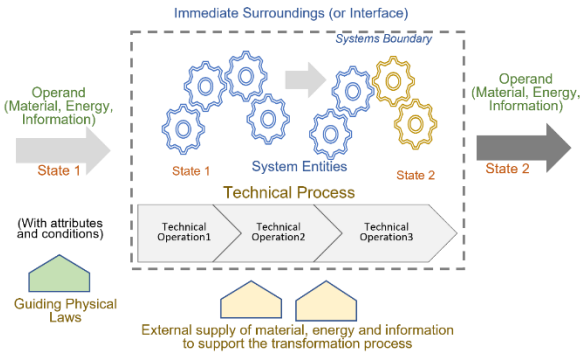


Figure 2. Categories of the extracted information

Table 3 shows the rule for information extraction for each category and the corresponding SAPPhIRE construct. It should be noted that the words identified from the natural language text and belonging to a particular category (first column of Table 3) may not directly fit into the definition of a SAPPhIRE construct and therefore, may need modification. E.g., in this sentence,' Due to the Seebeck effect, a net EMF is generated in a thermocouple circuit.', the Action Verb in 'EMF is generated' indicates a technical process inside the thermocouple system. In that technical process, diffusion of charge carriers due to the Seebeck effect leads to developing the potential difference. Hence, a true Phenomenon corresponding to the 'EMF is *generated*' is '*Diffusion* of charge carriers resulting in the generation of EMF.' An example of information extraction using rules is shown in Appendix B.

Table 3. Rules for information extraction

Extracted Information Category with mapping to the SAPPhIRE construct	**Extraction Rules**
Extracted Information Category: System Entities and Interfaces	• Nouns (subjects or objects) representing 'material' entities in the natural language description • These Nouns should represent a specific physical component and not the description of a physical component

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

Mapped SAPPhIRE construct name: Parts	• This physical component must belong to the system or its interface
Extracted Information Category: Attributes and Conditions *Mapped SAPPhIRE construct name*: Organs	• Adjectives of the Nouns representing a specific physical entity • Adverbs that describe a condition of an action • Preposition between two nouns or Conjunction between two verb clauses, which represents a condition for an event • These attributes or conditions should be necessary for the technical process to take place inside the system • These attributes or conditions should remain unchanged during the technical process
Extracted Information Category: Guiding Physical Law *Mapped SAPPhIRE construct name*: Effect	• Scientific equation • Scientific names stated as 'Laws', 'Principles', 'Equation' etc., called out in the text (e.g., Newton's Law of Motion, Maxwell Equation) • These equations or scientific terminologies should describe the physical principle (or law) that is guiding the technical process inside the system
Extracted Information Category: Technical process inside the system *Mapped SAPPhIRE construct name*: Phenomena	• The Action verbs associated with the Nouns (material entities) representing some actions • The actions are taking place inside of system boundary
Extracted Information Category: Effect on System Environment *Mapped SAPPhIRE construct name*: Action	• The Action verbs associated with the Nouns (material entities) representing some actions • The actions are taking place outside of the system boundary
Extracted Information Category: External supply to the system *Mapped SAPPhIRE construct name*: Input	• Subjects and Objects of Action verbs. These subjects/objects may represent materials or energy or information • If subjects/object indicates material entities, it should not have been considered as part of the system (i.e it is not a Part) • Subjects/Objects could be a verb, representing action as input • These entities (material, energy, information) should be crossing the system boundary
Extracted Information Category: State Change *Mapped SAPPhIRE construct name*: State Change	• Look for the Nouns that are connected to an Action Verb with a Preposition. State Change should be a Noun indicating a change in a physical property • Look for the Subjects and Objects associated with any other verb which indicates 'change' in the value of the noun • Words should represent an attribute undergoing change

5 RESULTS AND DISCUSSION

Technical documents are written in various linguistic styles with varying levels of detail for the content. Therefore the choice of a document becomes critical since the linguistic expressions in the document can be ambiguous and may create interpretational differences due to several factors, such as cultural differences, problem context, etc. (Halevy et al., 2009). It was found that designers generate solutions out of multiple analogies by breaking the main design problem into smaller sub-problems. Designers most often search online sources for inspiration to solve these smaller sub-problems (Vattam et al., 2013). Hence to validate our method, we used short online technical articles representing the cases designers typically use during ideation. Such short descriptions are usually about 'how a system works' and therefore are causal descriptions of systems working. In this paper, we validate the effectiveness of the newly developed method with four short descriptions of the working of a system - (a) a mechanical lock,

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

(b) an electrical battery, (c) a solar water heater and (d) visualization of infrared light by fish eye. The contents of these articles were hand curated with information taken from commonly available websites such as howstuffworks.com, explainthatstuff.com and asknature.org. We used the open source NLP software library in Python from spaCy (https://spacy.io/) to identify the POS tags for generating the knowledge graph. The rule based reasoning using Table 3 was, however, applied manually.

With the above four examples, we conducted an intrinsic evaluation of our method using the classical performance metrics, namely, precision, recall and F1-score (Moens, 2006). We calculated the performance metrics for each category of the extracted information (Table 4) and for the whole example (Table 5). In these four examples, the accuracy of extracted information (true vs. false) was determined by independent verification using systems modeling domain knowledge and input from SAPPhIRE modeling specialists.

Table 4. Category wise performance metrics

Labels	Precision	Recall	F1-score
Attributes and Conditions	97.0%	100.0%	98.5%
Effect on Environment	100.0%	100.0%	100.0%
External Input	92.3%	100.0%	96.0%
Guiding Physical Laws	100.0%	100.0%	100.0%
State Change	88.9%	80.0%	84.2%
System Entities	83.7%	90.0%	86.7%
Transformation Process	96.3%	92.9%	94.5%
Overall	91.9%	93.9%	92.9%

Table 5. Example wise performance metrics

Example ID	Precision	Recall	F1-Score
Example-1	97.1%	84.6%	90%
Example-2	90.3%	100.0%	95%
Example-3	100.0%	95.8%	98%
Example-4	70.8%	100.0%	83%

We find that System Entities and State Change have a lower F1 score compared to the others. Although it is easy to find Nouns that represent 'material entity,' many times, they are not valid system entities necessary to explain causality in the system, leading to more false positives among system entities. In the case of State Changes, natural language descriptions often are not very explicit about the changing attributes. Hence they remain undetected, leading to higher false negatives. Example-4 has the lowest F1 score among the four examples due to a higher proportion of false positive system entities. We shall include the lessons from these four examples in future improvement (refer to Appendix C).

6 CONCLUSIONS AND FUTURE WORK

Manually creating a causal representation from information given in technical documents requires significant effort by specialists. Machine learning techniques are used in automatic Information Extraction. In this paper, we presented an alternate method of information extraction using a Knowledge Graph and Rule-based reasoning approach. Results indicate this approach is promising, with its performance metrics being the same or higher in some cases than previously reported methods. Most importantly, the performance metrics reported here indicate the end-to-end process's performance since no pre- and post-processing tasks are involved.

This new approach has several advantages, as follows:

- There is no need for an extensive dataset to train and validate a model. We can use commonly used standard Language Models of NLP for POS tagging. These models are extensively validated with a very large corpus for the English language and are known to produce high accuracy
- The approach reported in this paper eliminates the effort intensive pre-processing steps before applying a machine learning model, as required in other reported methods
- We stayed close to the cognitive process of manually creating the SAPPhIRE models by specialists and thereby, we adopted an optimum approach to achieve consistency and accuracy
- The rule based reasoning method uses the domain knowledge of systems modeling. Therefore it helps to minimize inadvertent mistakes due to inadequate domain knowledge

However, we see the following limitations in the current state of the work as reported in this paper:

- Validation of the new method was carried out with four examples. Due to the complexity and variations of the natural language texts styles, extensive validation would be required

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

- The approach requires a manual effort to capture and document rules instead of automatically discovering them from the data
- Further improvement scope is identified for the rules (Appendix C). In future, as new rules are identified/developed over time, these need to be added to the rule base manually

We shall expand this work further to look into the improvement opportunities from this work. Our ultimate goal is to develop a generalized procedure to build a causal ontology by extracting information from technical documents. Therefore we shall look into characterizing various technical documents based on their source (e.g., knowledge articles on the web, research papers, patents, and design documents) to build a robust automation strategy. A trade-off study of different ontology-based information extraction methods, an experimental validation protocol with performance assessment criteria for the final method, and the benchmarking will be considered in the future course.

REFERENCES

Baldussu, A., Cascini, G., Rosa, F., & Rovida, E., (2012), "Causal models for bio-inspired design: a comparison", *DS 70: Proceedings of DESIGN 2012, the 12th International Design Conference*, Dubrovnik, Croatia (pp. 717-726).

Bhattacharya, K., Bhatt, A.N., Ranjan, B.S.C., Keshwani, S., Srinivasan, V. and Chakrabarti, A., (2023), "Extracting Information for Creating SAPPhIRE Model of Causality from Natural Language Descriptions", *Design Computing and Cognition'22* (pp. 3-20). https://doi.org/10.1007/978-3-031-20418-0_1

Cascini, G., & Zini, M., (2011), "Computer-aided comparison of thesauri extracted from complementary patent classes as a means to identify relevant field parameters", *Global Product Development* (pp. 555-566). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15973-2_56

Chakrabarti, A., Sarkar, P., Leelavathamma, B., & Nataraju, B. S., (2005), "A functional representation for aiding biomimetic and artificial inspiration of new ideas", *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 19(2), 113-132. https://doi.org/10.1017/s0890060405050109

Cheong, H. and Shu, L.H., (2012), "Automatic extraction of causally related functions from natural-language text for biomimetic design", In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 45066, pp. 373-382). American Society of Mechanical Engineers. https://doi.org/10.1115/detc2012-70732

Cowie, J. and Lehnert, W., (1996), "Information extraction", *Communications of the ACM*, 39(1), pp.80-91. https://doi.org/10.1145/234173.234209

Fu, K., Moreno, D., Yang, M. and Wood, K.L., (2014), "Bio-inspired design: an overview investigating open questions from the broader field of design-by-analogy", *Journal of Mechanical Design*, 136(11), p.111102. https://doi.org/10.1115/1.4028289

Gangemi, A., (2013), "A comparison of knowledge extraction tools for the semantic web", *Extended Semantic Web Conference* (pp. 351-366). Springer, Berlin. https://doi.org/10.1007/978-3-642-38288-8_24

Goel, A., Hagopian, K., Zhang, S., & Rugaber, S., (2022), "Towards a virtual librarian for biologically inspired design", *Design Computing and Cognition'20* (pp. 369-386). Springer, Cham. https://doi.org/10.1007/978-3-030-90625-2_21

Goel, A., Rugaber, S., & Vattam, S., (2009), "Structure, behavior & function of complex systems: The SBF modeling language", *Artificial Intelligence in Engineering Design, Analysis and Manufacturing*, 23(1), 23-35. https://doi.org/10.1017/s0890060409000080

Goel, A.K. and Shu, L.H., (2015), "Analogical thinking: An introduction in the context of design", *Artificial Intelligence in Engineering Design, Analysis and Manufacturing*, 29(2), pp.133-134. https://doi.org/10.1017/s0890060415000013

Gonçalves, M., Cardoso, C. and Badke-Schaub, P., (2016), "Inspiration choices that matter: the selection of external stimuli during ideation", *Design Science*, *2*. https://doi.org/10.1017/dsj.2016.10

Han, J., Sarica, S., Shi, F. and Luo, J., (2022), "Semantic Networks for Engineering Design: State of the Art and Future Directions", *Journal of Mechanical Design*, 144(2). https://doi.org/10.1115/1.4052148

Halevy, A., Norvig, P., & Pereira, F., (2009), "The Unreasonable Effectiveness of Data", *IEEE Intelligent Systems*, 24(2), 8–12. https://doi.org/10.1109/mis.2009.36

Howard, T.J., Dekoninck, E.A. and Culley, S.J., (2010), "The use of creative stimuli at early stages of industrial product innovation.", *Research in Engineering design*, 21(4), pp.263-274. https://doi.org/10.1007/s00163-010-0091-4

Hsieh, S.H., Lin, H.T., Chi, N.W., Chou, K.W. and Lin, K.Y., (2011), "Enabling the development of base domain ontology through extraction of knowledge from engineering domain handbooks", *Advanced Engineering Informatics*, 25(2), pp.288-296. https://doi.org/10.1016/j.aei.2010.08.004

Hubka, V. and Eder, W.E., (2002), "Theory of technical systems and engineering design synthesis", *Engineering design synthesis* (pp. 49-66). Springer, London. https://doi.org/10.1007/978-1-4471-3717-7_4

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

Keshwani, Sonal., (2018), *Supporting Designers in Generating Novel Ideas at the Conceptual Stage Using Analogies from the Biological Domain.*, IISc Bengaluru PhD Thesis (https://etd.iisc.ac.in/handle/2005/4205)

Li, Z. and Ramani, K., (2007), "Ontology-based design information extraction and retrieval", *Artificial Intelligence Engineering Design Analysis and Manufacturing*, 21(2), pp. 137-154. https://doi.org/10.1017/s0890060407070199

Moens, M.F., (2006), *Information extraction: algorithms and prospects in a retrieval context*, Heidelberg: Springer., 1–22. https://doi.org/10.1007/978-1-4020-4993-4_1

Nédellec, C., Nazarenko, A. and Bossy, R., (2009), "Information extraction", In *Handbook on ontologies* (pp. 663-685). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-92673-3_30

J Russell, S. and Norvig, P., (2010), *Artificial Intelligence A Modern Approach*, Third Edition, Prentice Hall

Sarica, S. and Luo, J., (2021), "Design knowledge representation with technology semantic network", *Proceedings of the Design Society, 1*, pp.1043-1052. https://doi.org/10.1017/pds.2021.104

Siddharth, L., Blessing, L., Wood, K.L. and Luo, J., (2022), "Engineering knowledge graph from patent database", *Journal of Computing and Information Science in Engineering*, 22(2). https://doi.org/10.1115/1.4052293

Tang, J., Hong, M., Zhang, D.L. and Li, J., (2008), "Information extraction: Methodologies and applications", *Emerging Technologies of Text Mining: Techniques and Applications* (pp. 1-33). IGI Global. https://doi.org/10.4018/978-1-59904-373-9.ch001

Vargas-Vera, M., Motta, E., Domingue, J., Shum, S.B. and Lanzoni, M., (2001), "Knowledge Extraction by Using an Ontology Based Annotation Tool", In *Semannot@ K-CAP* 2001.

Vattam, S. and Goel, A. (2013), "Seeking bioinspiration online: a descriptive account", *DS 75-7: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies*, Vol. 7: Human Behaviour in Design, Seoul, Korea, 19-22.08. 2013.

Venkataraman, S. and Chakrabarti, A., (2009), "Sapphire–an approach to analysis and synthesis", In *DS 58-2: Proceedings of ICED 09, the 17th International Conference on Engineering Design*, Vol. 2, Design Theory and Research Methodology, Palo Alto, CA, USA, 24.-27.08. 2009 (pp. 417-428).

Wimalasuriya, D.C. and Dou, D., (2010), "Ontology-based information extraction: An introduction and a survey of current approaches", *Journal of Information Science*, 36(3), pp.306-323. https://doi.org/10.1177/0165551509360123

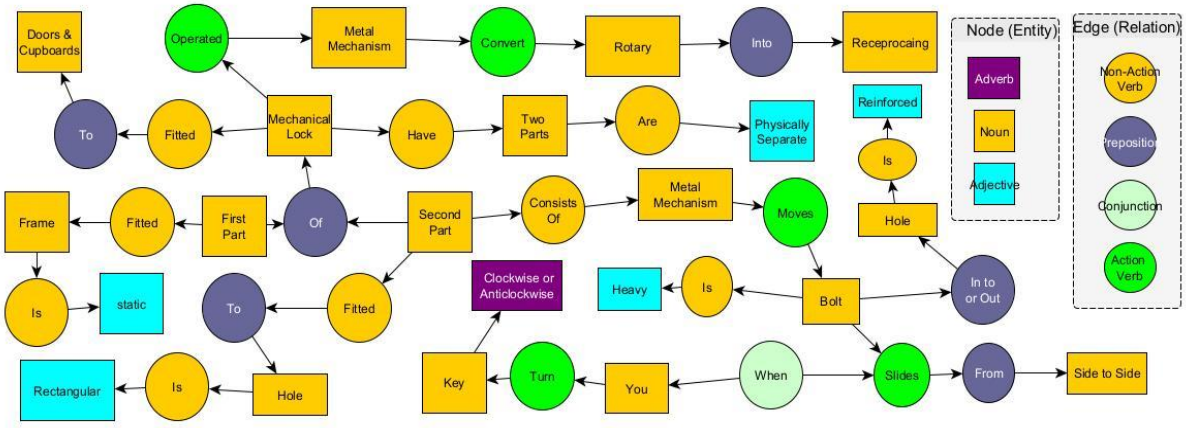
Yang, J., Kim, E., Hur, M., Cho, S., Han, M. and Seo, I., (2018), "Knowledge extraction and visualization of digital design process", *Expert Systems with Applications*, 92, pp. 206-215. https://doi.org/10.1016/j.eswa.2017.09.002

Zhang, C., Zhou, G., Chang, F. and Yang, X., (2020), "Learning domain ontologies from engineering documents for manufacturing knowledge reuse by a biologically inspired approach", *The International Journal of Advanced Manufacturing Technology*, 106(5), pp.2535-2551. https://doi.org/10.1007/s00170-019-04772-1

APPENDIX A

Sample text (*curated content with input from explainthatstuff.com*): "Mechanical-locks are fitted to doors and cupboards. Mechanical locks have two physically separate parts. First part is fitted to the frame. Frame is a static part. The second part of the lock fits into a rectangular hole. Second part consists of a metal-mechanism moving a heavy bolt into or out from the reinforced hole. The bolt slides from side to side when you turn a key in clockwise or anticlockwise direction, so the mechanical lock has to be operated by a mechanism to convert rotary motion of the key into reciprocating motion of the bolt."

Knowledge Graph representation of Mechanical Lock based on the above sample text:



https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press

APPENDIX B

Example of information extraction using rules.

Information Category	**Extracted Words**	**Remarks**
System Entities and Interfaces	Mechanical-Lock, Doors, Cupboards, First part, Frame, Second part, Hole, Metal-Mechanism, Heavy Bolt, Hole, Key	Doors and Cupboards are not system entities and do not take part in the system causality
Attributes and Conditions	*physically separate* parts, *Static* part, *Rectangular* hole, *Reinforced* hole, *Heavy* Bolt, Bolt slides *When* the key is turned	Except "when key is turned", other words indicate physical attributes of the system entities. The word "when" conveys a condition for the lock to work
Guiding Physical Law	None are given in this natural language text	
Technical process inside the system	*Turning* the Key, *Sliding* or *Moving* the bolt, *Operated* by a mechanism, *Convert* rotary motion to reciprocating motion	These verbs represent action inside a mechanical lock but do not represent any exchange of material, energy, or information. Hence, further post-processing is required to fit them into the definition of Phenomena
Effect on System Environment	None are given in this natural language text	
External supply to the system	You *turn* the key	Due to the turning of the key, motion (or energy) crosses the system boundary
State Change	Moving *into* or *out* / Slides *side* to *side*, Turning the key *in clockwise* or *anticlockwise*, Convert *rotary-motion* to *reciprocating-motion*	Though these do not directly say which attribute is changing, all these actions indicate associated state changes.

APPENDIX C

Areas for improvement in the rules for rules-based reasoning:

- Nouns used as an example or illustration and not a physical entity of the system under consideration
- A physical entity that doesn't take part in the causality of the system under consideration
- Nouns refer to variants of the system instead of being a component of the system
- An action verb implying a change but the document doesn't mention the attribute which is changing
- The attributes of the material, energy or information that cross the system boundary and are relevant for the system interactions

https://doi.org/10.1017/pds.2023.23 Published online by Cambridge University Press