Semantic Frame-Based Information Extraction from Utility Regulatory Documents to Support Compliance Checking

27

Xin Xu and Hubo Cai

Abstract

Computer-aided compliance checking is always a challenge in the Architecture, Engineering, and Construction (AEC) domain. In recent years, semantic compliance checking has gained a lot of attention. As the critical ingredient of the checking system, rule information needs to be extracted from regulatory texts and be formalized into machine-readable format. This paper proposed a semantic frame-based information extraction method with a focus on domain semantics and lexical semantics. The extraction process is characterized by the enrichment of lexical semantic frames and the mapping with the domain semantic framework. Natural language processing (NLP) and machine learning (ML) techniques were used to implement the extraction mechanism. The preliminary experiment shows a promising performance when extracting rule information from Indian Utility Accommodation Policy.

Keywords

Compliance checking • Semantic framework • Semantic frame • Natural language processing (NLP)

27.1 Introduction

Research of computer-aided compliance checking is always a challenge, which has been studied and developed over the last four decades, with a focus on buildings and the built environment [1]. In recent years, computational implementation and tools (e.g., Solibri Model Checker, Jotne EDModelChecker, FORNAX, and SMARTcodes) have been developed by practitioners and software developers, adopting a variety of approaches including the widely popular rule-based systems [2]. One of the most important steps in a rule-based compliance checking system is the rule requirement analysis, which targets at the extraction of rule knowledge from regulatory documents. However, most current initiatives rely on manual efforts to extract requirements from regulatory documents and encode these requirements in a machine-interpretable format, which is time-consuming, costly and error prone [3]. Moreover, rule knowledge is conventionally represented in natural language texts and formalized in various formatting and semantic structures. Even in a single regulation, the formatting and semantics of the provisions could vary from one chapter to another. Consequently, the task of information extraction from regulatory documents is very complicated. In the Architecture, Engineering and Construction (AEC) domain, several approaches that aim at interpreting regulatory documents, extracting rule knowledge, and representing rules in a standard format have been attempted, including the use of document markup techniques to aid navigating the document structure and the integration of natural language processing (NLP) algorithms and semantic web technologies to facilitate syntactic and semantic interpretations of regulatory sentences [3–7].

Since more and more design and construction data is represented in the Resource Description Framework (RDF) data model [8], the underlying semantic and logical basis can provide an effective platform for implementing semantic compliance checking, as proposed by Pauwels et al. [9]. To ensure an effective semantic checking performance, the extracted rule

X. Xu · H. Cai (✉)
Purdue University, West Lafayette, IN 47907, USA
e-mail: hubocai@purdue.edu

I. Mutis and T. Hartmann (eds.), *Advances in Informatics and Computing in Civil and Construction Engineering*,
https://doi.org/10.1007/978-3-030-00220-6_27

information needs to follow a formalized representation format that aligns well with RDF-based design and construction data model. To achieve this goal, this paper proposed a novel approach to extract the rule information from regulatory documents by using NLP techniques and integrate domain-level semantic framework and sentence-level semantic frames. The application is targeted at checking the underground utility spatial configuration against the utility accommodation policy to identify potential utility conflicts.

27.2 Methodology

Substantial efforts on information extraction exist outside the AEC domain [10–12]. They focus on named entity extraction, attribute extraction, relation extraction, and event extraction [13]. Rule-based and machine learning-based approaches are frequently used for text processing. Semantic approach (using meaning/context-related features captured by a domain ontology) is expected to enhance the information extraction performance over the existing approaches [10]. In the AEC domain, similar research efforts, especially the ontology-based information extraction, are very limited. To take a step further, the proposed methodology in this paper used sentence-level semantic frames to boost the rule information extraction together with a domain semantic framework (a domain ontology).

27.2.1 Semantic Framework and Semantic Frames

In urban infrastructure domain, relevant work has been untaken to develop an ontology for products [14], processes [15], and actors [16]. Salama and El-Gohary [6] proposed a semantic framework for the compliance checking knowledge in the construction domain. By combining the existing ontological models, a semantic framework in the context of utility compliance checking was developed. A partial view of the semantic framework (ontology) is shown in Fig. 27.1. The development process is outside of the scope of this paper and is explained in Xu and Cai [17].

The developed ontology includes concepts applicable in the context of utility compliance checking and categorized into four main top-level concepts, namely, Utility Element, Surrounding Element, Rule Element, and Compliance Checking Element. This ontology also remains flexible and extendable. Most of the concepts and instances observed in this application domain can be organized and represented within this semantic framework. More efficient rule information extraction algorithms can be devised under the guidance of this framework as discussed in 27.2.3.

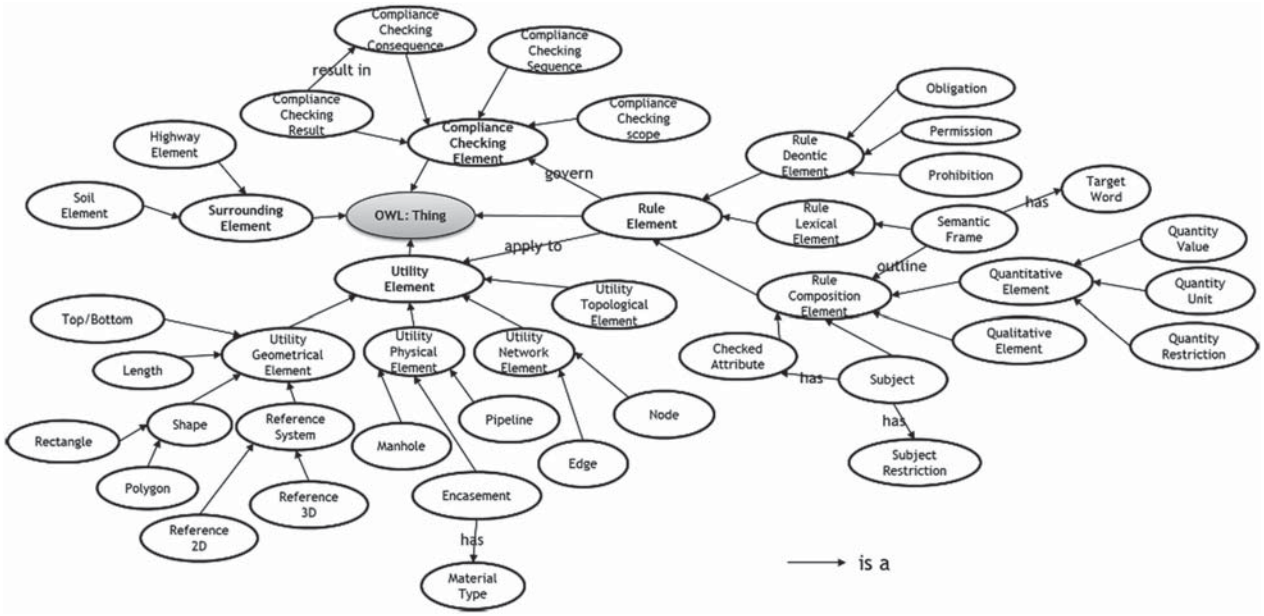

Fig. 27.1 Partial view of the ontology

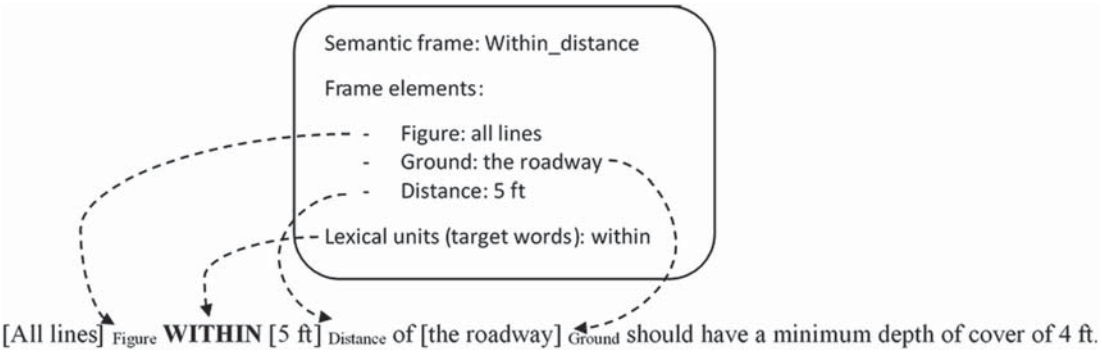

Fig. 27.2 Semantic frame filled with information extracted from rule clause

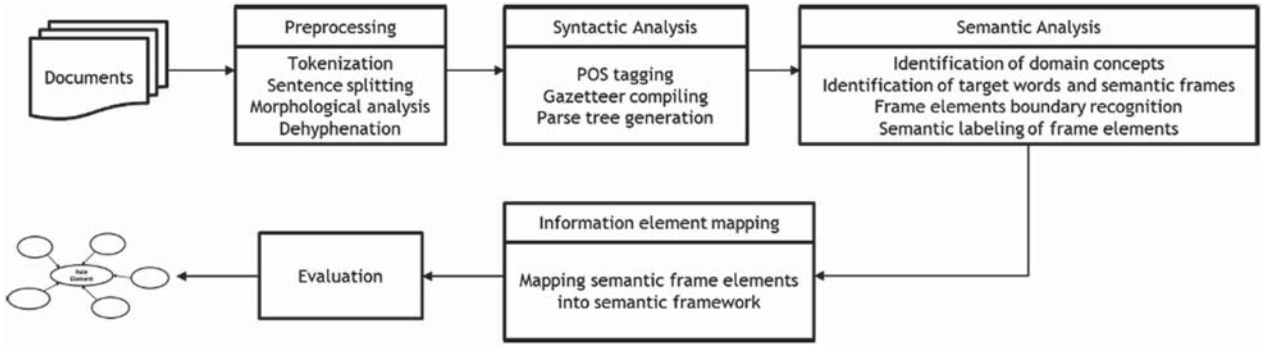

Fig. 27.3 Proposed information extraction methodology

Unlike other rule information extraction methods, this paper also added sentence-level semantic frames into the semantic framework to capture rule lexical elements. Semantic frame is a description of a type of event, relation, or entity and the participants in it, on basis of which sentences can be best understood [18]. Within a semantic frame, several core frame elements exist to contain important information entities extracted from sentences. Words that evoke the frame are called lexical units or target words. An excerpt from Indiana utility accommodation policy, all lines within 5 ft of the roadway should have a minimum depth of cover of 4 ft, can be partially interpreted via the within_distance semantic frame as in Fig. 27.2.

The readers are referred to the FrameNet project (https://framenet.icsi.berkeley.edu/fndrupal/) for detailed explanations. FrameNet has begun to annotate some continuous texts, as a demonstration of how frame semantics can contribute to text understanding and this style of annotation typically involves marking frame elements of frames evoked by target words in each sentence. Therefore, FrameNet can provide sufficient annotated training data for extracting frame elements from rule sentences. To this end, the procedure of semantic frame-based information extraction is summarized in Fig. 27.3.

27.2.2 Preprocessing

This phase is used to prepare the raw texts from regulatory documents for further processing. The rule sentences selected from utility accommodation polies were preprocessed following the procedure of tokenization, sentence splitting, morphological analysis, and de-hyphenation.

Tokenization is the process of splitting the raw texts into tokens, where a token is a word, a number, a symbol, or a whitespace [19]. Sentence splitting aims at detecting sentence boundary indicators (i.e., periods, exclamation marks, and question marks) and recognizing each sentence of the texts [13]. Morphological analysis aims to map the different forms of a word (e.g., plural form of a noun) to its lexical form (e.g., singular form of a noun) [20]. De-hyphenation is to remove hyphens that are used to continue a word across two lines. To illustrate, the same excerpt, all lines within 5 ft of the roadway should have a minimum depth of cover of 4 ft, was preprocessed as shown in Fig. 27.4.

Fig. 27.4 Illustrative example of preprocessing

Original text:
All lines within 5 ft of the roadway should have a minimum depth of cover of 4 ft.
Preprocessed text:
<sentence>
<Token>all</Token><Token>line</Token><Token>within</Token>
<Token>5</Token><Token>foot</Token><Token>of</Token><Token>the</Token>
<Token>roadway</Token><Token>should</Token><Token>have</Token>
<Token>a</Token><Token>minimum</Token><Token>depth</Token>
<Token>of</Token><Token>cover</Token><Token>of</Token><Token>4</Token>
<Token>foot</Token>
</sentence>

27.2.3 Syntactic Analysis

This phase is used to analyze the syntactic features of the preprocessed texts, which consists of part-of-speech (POS) tagging, gazetteer compiling, and parse tree generation.

POS tagging is to assign parts of speech to each word based on its syntactic word category, such as NN (singular nouns), DT (determiner), JJ (adjectives), VB (verb), and CD (cardinal number). The readers are referred to the Penn Treebank Project (https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) for the complete list of POS tags. Figure 27.5 demonstrates the tagged result of the same excerpt using Stanford Parser.

Gazetteer compiling aims to group a set of terms based on any specific commonality possessed by these terms. By investigating the regulatory texts in the application domain, the negation gazetteer list (e.g., no, not), comparative relation gazetteer list (e.g., less than, greater or equal, minimum), and unit gazetteer list (e.g., foot, meter) were compiled, as similar as proposed by Zhang and El-Gohary [3]. If a word or a phrase is found within the compiled gazetteer list, a specific tag will

POS tagged text:
All/DT lines/NNS within/IN 5/CD ft/NN of/IN the/DT roadway/NN should/MD have/VB a/DT minimum/JJ depth/NN of/IN cover/NN of/IN 4/CD ft/NN.

Fig. 27.5 Illustrative example of POS tagging

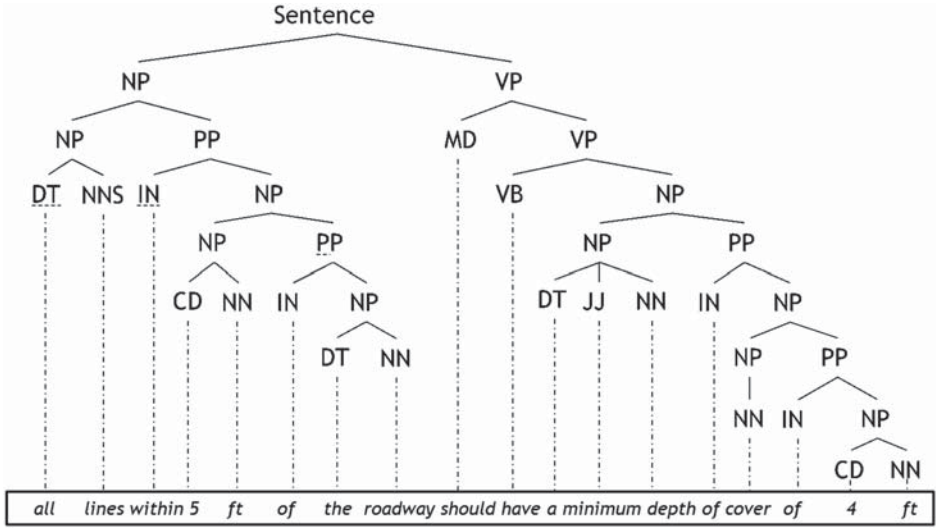

Fig. 27.6 Illustrative example of generated parse tree

be assigned. For example, in the POS tagged text, "minimum" was detected by gazetteer lookup then was labelled as CR (comparative relation gazetteer). The gazetteer tags together with POS tags can assist in subsequent information extraction phases, which are discussed in the following sections.

Parse tree is an ordered, rooted tree that represents the syntactic structure of a sentence according to some context-free grammar (CFG), starting from Sentence and ending in POS tagged words. The parse tree of the same excerpt was generated using Stanford Parser, as illustrated in Fig. 27.6.

The terminals (POS tags) are chunked into phrasal tags such as NP (noun phrase), VP (verb phrase) and PP (prepositional phrase) based on CFG following the bottom to up direction. Once rule clauses are decomposed into parse tree, rule information entities can then be extracted on the basis of their phrasal tags. In addition, the tree path (up or down) would be very helpful in assigning semantic roles to frame elements, which are discussed in 27.2.4.

27.2.4 Semantic Analysis

This phase is used to analyze the semantic features of the syntactically processed texts based on domain-level semantics (ontology) and lexical semantics (semantic frames). The goal is to identify the domain concepts structured in the semantic framework and to fill the semantic frames evoked by target words that are detected in the texts.

Identification of domain concepts. Since the ontology captures the concepts related in the context of utility compliance checking, it can be regarded as a gazetteer list populated with domain concepts, thus assisting the extraction of relevant information with domain-specific meanings from the texts. In previous work, the authors proposed an ontology to capture the concepts about utility physical products and attributes, which can be incorporated under the concept of utility physical element. For instance, the "pipeline" concept is a "utility physical element" concept. The "depth of cover" concept is a utility attribute concept that can be linked with the "pipeline" concept using the "has" relationship. Using these concepts modeled in the ontology, "lines" and "depth of cover" can be extracted from the same excerpt, "all lines within 5 ft of the roadway should have a minimum depth of cover of 4 ft", by matching with the concepts of "pipeline" and "depth of cover" respectively. By this way, domain semantic tags can be added into the syntactically processed texts. Moreover, cross-concept relationships like the "is a" relationship and the "has" relationship defined within the ontology can aid in the semantic labeling of frame elements, which is discussed in the following sections.

Identification of target words and semantic frames. In the proposed methodology, lexical semantics was incorporated into the semantic framework by analyzing the target words and the evoked semantic frames. Given a sentence, once the target words are identified and the corresponding semantic frames are filled with core frame elements, the sentence can be interpreted from the lexical perspective. Information entities and relations can also be extracted based on the filled semantic frames. Following this direction, the authors investigated the utility accommodation policies of six states (Indiana, Tennessee, Illinois, Ohio, Michigan, and Kansas) to observe the rule clause patterns. With reference to semantic frames built in the FrameNet, three applicable semantic frames to the utility accommodation policy sentences are summarized in Fig. 27.7. Core frame elements and possible target words are also listed in Fig. 27.7. One observation was found in the preliminary investigation that a single rule sentence may hold multiple semantic frames associated with multiple target words. In order to extract all rule information entities and relations, all applicable semantic frames must be filled. In 27.2.1, the excerpt was interpreted using the semantic frame of "within_distance" evoked by the target word of "within". What is still missing is the "deontic_rule" semantic frame evoked by the target word of "should". Only by enriching these two semantic frames can all the rule knowledge within the excerpt be extracted.

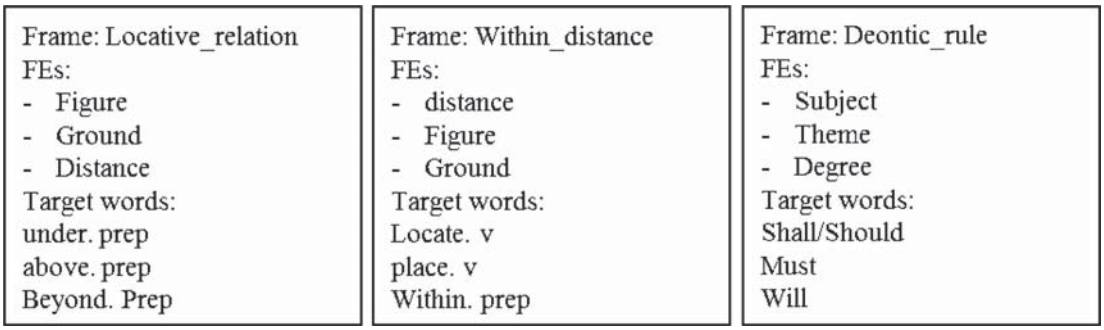

Fig. 27.7 Examples of applicable semantic frames

Analysis of lexical semantics starts with the identification of the target words, then the selection of applicable semantic frames, and lastly the enrichment of the semantic frames. To identify the target words, the highest ranked word in the sentence is selected by ranking all sentence words by semantic similarity to the evoking words listed in FrameNet [21]. Semantic similarity is computed following the same procedure as reported in [22]. In the case of this paper, the authors enumerated all the semantic frames with all the possible target words formalized in a lookup table. Then the target word in a given sentence can be found by simply referring to the lookup table. Correspondingly, evoked semantic frames can be matched. The following procedure is to fill the applicable semantic frames with frame elements, which is discussed in next two sections.

Recognition of frame elements boundary. After generating a set of syntactic features (e.g., POS tags, gazetteer tags, and phrasal tags) and domain semantic features (domain semantic tags), some frame elements boundary can be easily detected based on these features. For instance, domain semantic tags and gazetteer tags are often assigned to frame elements. POS tags following a specific sequence can also form a frame element, which plays a similar role as phrasal tags. Quite a few studies used regular expressions to encode the sequence patterns of POS tags [3, 7, 23]. The detailed explanation for the syntax of regular expressions is given in [24]. The NP tag corresponds to a list of POS tags matching the pattern (<DT|CD>? <JJ>*<NNS>). The phrases in the excerpt, *all lines* and *the roadway* are examples that satisfy the pattern and they are both frame elements in the "within_distance" frame. Another way of detecting the boundary of frame elements is to check the relative location with the target word, which can be analyzed along the generated sentence parse tree. The search path can be expressed as a sequence of non-terminal nodes of the parse tree linked by direction symbols (up or down). For instance, search along the path of "within" ↑ PP ↓ NP ↓ NP for the frame element of "5 ft". When the sentence structure gets much complicated, a collection of sample sentences is required to develop a set of reliable heuristics for targeting the boundary of those frame elements.

Semantic labeling of frame elements. As described previously, multiple core frame elements exist in a semantic frame. In the "within_distance" frame, "figure", "ground", and "degree" are three semantic roles labeled to the frame elements (see Fig. 27.2). Once the frame elements are recognized, semantic role labels are required to assigned to them, which is also critical to interpret their relations.

Research on semantic role labeling has been studied for long in the area of computational linguistics [21, 25]. Recently, a number of researchers have used machine learning techniques to build system which can be trained on FrameNet annotation data (as described in 27.2.1) and automatically produce similar annotation on new (previously unseen) texts [25, 26]. This process is called automatic semantic role labeling. In the case of this paper, the task of semantic role labeling is to classify the recognized frame elements into their corresponding semantic roles. The authors used a collection of annotated data from FrameNet to train a probability model that can be relied on to annotate new data from utility rule clauses. Features that are used in the probability model are summarized in Table 27.1. Some features were first designed in [27]. The examples in Table 27.1 all refer to the same excerpt from utility accommodation policy and some features are generated from the syntactic processing phase.

Table 27.1 Feature set

Feature type	Description
Phrase type	This feature indicates the syntactic type of the phrase recognized as a frame element, e.g., NP for "all lines" labeled as the role of "Figure"
Parse tree path	This feature contains the path in the parse tree between the recognized frame element and the target word, expressed as a sequence of nonterminal labels linked by direction symbols (up or down), e.g., NP ↑ NP ↓ PP for "all lines" labeled as the role of "Figure"
Position	This feature indicates if the frame element appears before or after the target word in the processed sentence, e.g., "all lines" appears before the target word "within"
Voice	This feature distinguishes between active or passive voice for the target word
Target word	This feature indicates the target word identified in the sentence with the case and morphological information preserved
Domain concept class	This feature indicates the concept class that the recognized frame element belongs to, e.g., "all lines" belongs to the concept class of pipeline in the semantic framework
Domain relationship class	This feature indicates the possible domain relationship the recognized frame element holds to the target word or other frame elements, e.g., "all lines" and "depth of cover" may be linked with a "has" relationship in the semantic framework
Gazetteer class	This feature indicates the prebuilt gazetteer class that the recognized frame element belongs to

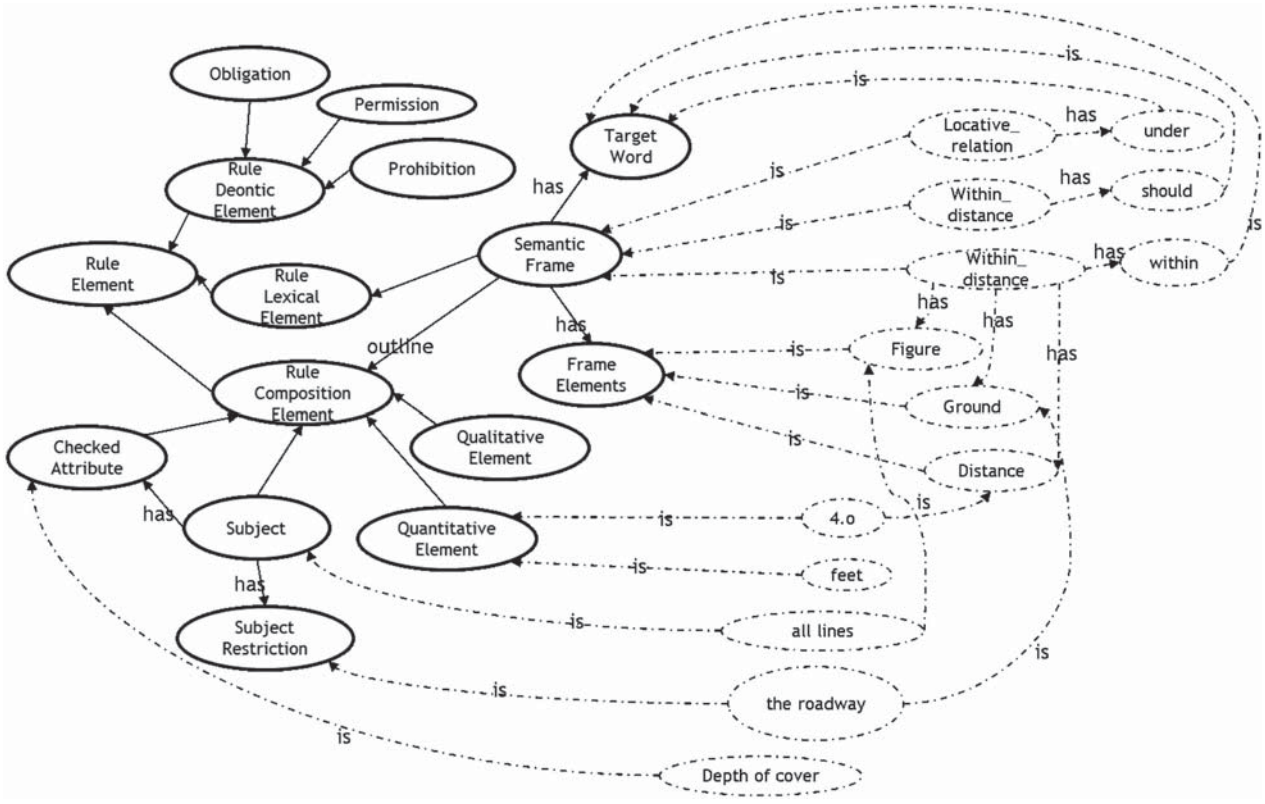

Fig. 27.8 Partial view of the mapping result

In FrameNet, a semantic frame has also a description that defined the relations holding between its frame elements, which is called the scene of the frame. For example, the scene of the "within_distance" frame is defined as: the "Figure" holds a "within" spatial relationship with the "Ground" and the distance between them is the "Distance". Once all the frame elements are labeled with semantic roles using the machine leaning-based system, the semantic relationships between frame element can be extracted correspondingly.

27.2.5 Information Element Mapping

The last step of the semantic frame-based information extraction is to map the semantic frame elements into the semantic framework in a unified knowledge representation format, which is required by semantic-based checking mechanism. The mapping process can be regarded as the process of adding instance data extracted from regulatory texts into the semantic schema. Figure 27.8 gives an example for the mapping result of processing the same excerpt (solid circles hold the semantic concepts and dash circles hold the instance data). Different semantic frames may have different mapping rules. The mapping can be done through comparing lexical semantics with domain semantics and linking them with predefined domain relationships.

27.3 Conclusion

As described previously, semantic frame-based information extraction is more suitable for semantic-based compliance checking. In the context of semantic checking, either fact knowledge or rule knowledge is represented in RDF graph. An RDF graph is constructed by applying a logical AND operator to a series of logical statements containing concepts and their relationships, similar as Figs. 27.1 and 27.8. These logical statements are often referred to RDF triples, consisting of a subject, a predicate, and an object. It is observed that the structure of the semantic frames displays a Subject-Verb-Object

dependency with the target word. Given one sentence, all Subject-Verb-Object triples can be collected by identifying all the frame elements of all applicable semantic frames. Therefore, semantic frame-based information extraction serves as a starting point for semantic-based compliance checking, which gains a lot of popularity in recent years in AEC domain [2, 9].

The authors' preliminary experiment demonstrates the promising performance of the semantic frame-based information extraction method. The authors tested the method in extracting rule information from Indiana Utility Accommodation Policy using the semantic frame of "within_distance" with the target word of "within". The extracted rule information is concerned about the spatial relation between utility assets and their surrounding infrastructure including their relative distance. The precision is 92.32%. It can be expected that when more semantic frames are involved, the performance will deteriorate. Therefore, future experiments of rule information extraction using multiple semantic frames simultaneously are needed to improve the proposed methodology.

References

1. Dimyadi, J., et al.: Computerizing regulatory knowledge for building engineering design. J. Comput. Civil Eng. **30**(5) (2016)
2. Eastman, C., et al.: Automatic rule-based checking of building designs. Autom. constr. **18**(8), 1011–1033 (2009)
3. Zhang, J., El-Gohary, N.M.: Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. J. Comput. Civil Eng. **30**(2) (2013)
4. Abuzir, Y., Abuzir, M.D.O.: Constructing the civil engineering thesaurus (CET) using ThesWB. In: Computing in Civil Engineering, pp. 400–412 (2002)
5. Al Qady, M., Kandil, A.: Concept relation extraction from construction documents using natural language processing. J. Const. Eng. Manage. **136**(3), 294–302 (2009)
6. Salama, D.A., El-Gohary, N.M.: Automated compliance checking of construction operation plans using a deontology for the construction domain. J. Comput. Civil Eng. **27**(6), 681–698 (2013)
7. Zhou, P., El-Gohary, N.: Ontology-based automated information extraction from building energy conservation codes. Autom. Const. **74**, 103–117 (2017)
8. Manola, F., Miller, E., McBride, B.: RDF primer. W3C recommendation (2004)
9. Pauwels, P., et al.: A semantic rule checking environment for building performance checking. Autom. Const. **20**(5), 506–518 (2016)
10. Soysal, E., Cicekli, I., Baykal, N.: Design and evaluation of an ontology based information extraction system for radiological reports. Comput. Biol. Med. **40**(11–12), 900–911 (2010)
11. Sapkota, K., et al.: Extracting meaningful entities from regulatory text: Towards automating regulatory compliance. In: Requirements Engineering and Law (RELAW), 2012 Fifth International Workshop on IEEE, pp. 29–32 (2012)
12. Hogenboom, A., et al.: Semantics-based information extraction for detecting economic events. Multimed. Tools Appl. **64**(1), 27–52 (2013)
13. Jurafsky, D., Martin, J.H.: Speech and language processing. Pearson, London (2014)
14. El-Diraby, T.E., Osman, H.: A domain ontology for construction concepts in urban infrastructure products. Autom. Const. **20**(8), 1120–1132 (2011)
15. El-Gohary, N.M., El-Diraby, T.E.: Domain ontology for processes in infrastructure and construction. J. Const. Eng. Manage. **136**(7), 730–744 (2010)
16. Zhang, J., El-Diraby, T.E.: Mapping actors and roles in construction knowledge. In: Construction Research Congress 2009: Building a Sustainable Future, pp. 936–945 (2009)
17. Xu, X., Cai, H., Chen, K.: An ontology approach to utility knowledge representation. Const. Res. Congr. **2018**, 311–321 (2018)
18. Fillmore, C.J., Baker, C.: A frames approach to semantic analysis. The Oxford handbook of linguistic analysis (2010)
19. Moens, M.F.: Information extraction: algorithms and prospects in a retrieval context. Springer Science & Business Media (2006)
20. Fautsch, C., Savoy, J.: Algorithmic stemmers or morphological analysis? An evaluation. J. Assoc. Inf. Sci. Technol. **60**(8), 1616–1624 (2009)
21. Moschitti, A., Morarescu, P., Harabagiu, S.M.: Open Domain Information Extraction via Automatic Semantic Labeling. In: FLAIRS conference, pp. 397–401 (2003)
22. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. Advances in automatic text summarization, pp. 111–121 (1999)
23. Li, S., Cai, H., Kamat, V.R.: Integrating natural language processing and spatial reasoning for utility compliance checking. J. Const. Eng. Manage. **142**(12) (2016)
24. Thompson, K.: Programming techniques: regular expression search algorithm. Commun. ACM **11**(6), 419–422 (1968)
25. Das, D., et al.: Frame-semantic parsing. Comput. Linguist. **40**(1), 9–56 (2014)
26. Roth, M., Lapata, M.: Context-aware frame-semantic role labeling. Trans. Assoc. Comput. Linguist. **3**, 449–460 (2015)
27. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Comput. linguist. **28**(3), 245–288 (2002)