Leveraging Word Embeddings and Transformers to Extract Semantics from Building Regulations Text

Odinakachukwu Okonkwo[1,†], Amna Dridi[1,*,†] and Edlira Vakaj[1,†]

[1]Faculty of Computing, Engineering and Built Environment, Birmingham City University, B4 7XG, Birmingham, UK

Abstract

In the recent years, the interest to knowledge extraction in the architecture, engineering and construction (AEC) domain has grown dramatically. Along with the advances in the AEC domain, a massive amount of data is collected from sensors, project management software, drones and 3D scanning. However, the construction regulatory knowledge has maintained primarily in the form of unstructured text. Natural Language Processing (NLP) has been recently introduced to the construction industry to extract underlying knowledge from unstructured data. For instance, NLP can be used to extract key information from construction contracts and specifications, identify potential risks, and automate compliance checking. It is considered impractical for construction engineers and stakeholders to author formal, accurate, and structured building regulatory rules. However, previous efforts on extracting knowledge from unstructured text in AEC domain have mainly focused on basic concepts and hierarchies for ontology engineering using traditional NLP techniques, rather than deeply digging in the nature of the used NLP techniques and their abilities to capture semantics from the building regulations text. In this context, this paper focuses on the development of a semantic-based testing approach that studies the performance of modern NLP techniques, namely word embeddings and transformers, on extracting semantic regularities within the building regulatory text. Specifically, this paper studies the ability of *word2vec*, *BERT*, and *Sentence BERT (SBERT)* to extract semantic regularities from the British building regulations at both word and sentence levels. The UK building regulations code has been used as a dataset. The ground truth of semantic regulations has been manually curated from the well-established Brick Ontology to test the performance of the proposed NLP techniques to capture the semantic regularities from the building regulatory text. Both quantitative and qualitative analyses have been performed, and the obtained results show that modern NLP techniques can reliably capture semantic regularities from the building regulations text at both word and sentence levels, with an accuracy that reaches 80% at the word-level, and hits 100% at the sentence-level.

Keywords

AEC domain, Natural Language Processing, BERT, word2vec, Sentence BERT, Word embeddings, Transformers, Building regulations, Semantic regulations, Machine Learning,

1. Introduction

The automation of activities, which were done manually has become quite prevalent in our world. Textual data is now automated to extract interesting insights and information, and save time and man power. One such widely growing domain that fosters this digitisation process is

Proceedings LDAC2023 – 11th Linked Data in Architecture and Construction, June 15–16, 2023, Matera, Italy

*Corresponding author.

odinakachukwu.okonkwo@mail.bcu.ac.uk (O. Okonkwo); amna.dridi@bcu.ac.uk (A. Dridi); edlira.vakaj@bcu.ac.uk (E. Vakaj)
0000-0002-0185-103X (A. Dridi); 0000-0002-9421-8566 (E. Vakaj)

the Architecture, Engineering and Construction (AEC) domain. Natural Language processing (NLP) - as a branch of Machine learning (ML) dealing with textual data - has recently skyrocketed in the AEC domain, essentially with the success of word embeddings [1] and transformers [2]. Word embeddings [1] are numerical representations of words in a high-dimensional vector space. Transformers, on the other hand, are a type of neural network architecture for processing sequences of text [2]. Together, they have enabled significant advances in a wide range of NLP applications, among which applications in AEC domain such as safety management [3], automation compliance checking [4, 5, 6, 7, 8, 9, 10], public opinion analysis [11], building design [12, 13], contract management [14, 15], and others [16, 17, 18].

Due to their ability to reduce the gap between human and computer language comprehension, typically, word embeddings and transformers have been used as features for different ML tasks dealing with diverse aspects in AEC domain, such as text classification [19] and clustering [20]. Despite the sensitivity of the hyper-parameters of these NLP techniques and their characteristics of being data and task dependant [21], there are a few studies that deeply investigate the capabilities of word embeddings and transformers to capture the semantic regularities within a domain-specific text [22]. While it is important in the construction domain, specifically with regulatory text, to extract the required and exact information, an in-depth analysis of these NLP techniques with application to the building regulations text is becoming an urge.

This paper looks at the ability of word embeddings and transformers to extract semantic meaning from the building regulatory text. The term semantics became important because the information in a phrase or a piece of text is stored in organised sequences, with the semantic arrangement of words expressing the meaning of the text. This also implies that the integrity of the semantic meaning in the sentences must be maintained during the extraction of text. To this end, *word2vec* [23], *BERT* [2], and *Sentence BERT (SBERT)* [24] have been used as word embeddings and transformers techniques, respectively to test their abilities to capture semantics from the regulatory text in AEC domain. To make our point, we propose training these models with the UK building regulations code; moreover, we propose using common-sense knowledge manually curated from a well-established ontology in the building environment domain, namely *Brick Ontology*[1] [25]. As a result, this work adds breadth to the debate on the strengths of using modern NLP techniques for knowledge extraction from regulatory text in the construction industry. Although transformers have been widely used in similar works[5, 7, 8, 26], however, none of the existing work has studied their suitability for the task and how effective they are to capture the semantic regularities within the domain-specific use. This is important especially when it comes to critical downstream tasks like information extraction and rule generation. With regards to word embeddings literature, many researchers have studied the capabilities of these techniques to capture the semantic regularities with a domain-specific language, such as the medical domain [27] or the scientific domain [22]. However, to the best of our knowledge, no existing work has been done so far for the AEC-related language. Additionally, going beyond the aforementioned literature, this work studies also the suitability of transformers to capture the semantic regularities at both word and sentence-levels. To the best of our knowledge, the proposed work represents the first attempt to methodically test the ability of word embeddings and transformers to capture semantics in the building text.

[1] https://brickschema.org/ontology

We list the major contributions of this work as follows: (i) we propose the accuracy of word2vec, BERT and SBERT to capture the semantic regularities within the building text as an objective to measure while learning the models, (ii) we create an analogy dataset for the building regulations text by manually curating the Brick Ontology, and (iii) we evaluate our work quantitatively and qualitatively on a corpus generated from the UK building regulations code at both word and sentence levels. Our embeddings detected interesting semantic relations in AEC domain such as *"meter is to electricity as consumer_unit is to consumer"*, and *"room is a type of space as door is a type of fitting"*. The obtained results are, therefore, both promising and insightful.

The rest of the paper is organised as follows. Section 2 summarises the existing approaches on NLP in the construction industry and gives an overview on work that attempted to use word embeddings and transformers in the AEC domain. Section 3 presents our methodology and describes the proposed word embeddings and transformers techniques. Section 4 describes the dataset we have created from the UK building regulations code, the analogy dataset we have created from the Brick Ontology as gold standard, presents and discusses the obtained results. Finally, in Section 5 we conclude and draw future directions.

2. Related Work

This section summarises the literature on both NLP in AEC domain and semantic search with word embeddings and transformers, hence covering the two topics of this paper.

2.1. NLP in AEC domain

Zhang and El-Gohary were among the first researchers who applied NLP for automated information Extraction (IE) in AEC domain. They used a set of pattern matching based IE rules utilising a series of syntactic and semantic text features in the patterns of the building rules. They also utilised an ontology to support the identification of semantic text features. The IE algorithms built was tested in extracting quantitative requirements from the 2009 international building code and the results were 0.969 and 0.944 precision and recall, respectively. However, they opined that the use of Machine Learning algorithms for text processing yielded less precision and recall results when compared manually coded rules, which requires more human effort [5]. However, because the manual process lacked neither flexibility nor scalability; Whenever the building rules change, there will be a need to make adjustments to the building code, Zhang and El-Gohary [26], therefore, proposed the use of a deep neural networks for semantic and syntactic IE aspects from AEC regulation papers. The suggested approach performed well with an average accuracy of 93% and a recall of 92.9%[26]. In another work, Zhang *et al.*[10] used construction scene graphs and the C-BERT network, to propose an autonomous technique for hazard inference. Initially, computer vision was used to produce construction scene graphs with interaction-level scene descriptions that included entities, characteristics, and their interactions. Second, the C-BERT network was meant to infer potential dangers by combining scene graphs with domain information such as building rules. Five separate working settings were employed to illustrate the validity of the suggested method, which achieved an identification accuracy of

97.82%. It offered an effective mechanism for merging visual information and domain knowledge for automated safety monitoring and paved the way for huge multi-modal information fusion inside the industry.

In the same context of rule automation, Zhou *et al.* [28] proposed an automated rule interpretation system for automated compliance checking that interprets sentences into single requirement and multi-requirement rules. The parsing accuracy for basic sentences was 99.6%, exceeding the state of the art, and 91.0% parsing precision for complicated sentences, which are challenging for present algorithms to handle.

2.2. Word Embeddings and Transformers for Semantic Search in AEC domain

A.J.P. Tixier *et al.* [29] applied word embedding techniques(word2vec) on an 11-million word corpus obtained from the construction domain and obtained word vectors from the process. They explored the embedding space created by the vectors and affirmed that the word vectors were able to capture meaningful semantics related to construction specific concepts. They evaluated the performance of the vectors against the ones that were trained on a 100B-word corpus (Google News) within the confines of an injury report classification task and without any parameter tuning, their embeddings gave competitive results, and outperformed the Google News vectors in many cases.

Yuan *et al.* [30] devised a technique for determining phrase similarity based on the BERT model, and compared the classic ALBERT, ESIM, and BIMPM models. Their experimental findings demonstrated that the BERT model calculates text similarity with an accuracy of 87%, which is clearly superior to other models. Simultaneously, the synonym model is trained using Word2Vec to extract target-word-related synonyms.

Risch and Krestel[31] applied domain-specific word embedding techniques for the automation of patent applications. Here, they compared novelty applications for patents to existing patents in the same class. However, one challenge the task faced was patent-specific language use, especially in phrases and vocabulary. To account for this language usage, the authors proposed pre-trained word embeddings for the patent domain that are domain-specific. The authors trained the model on a massive dataset including more than 5 million patents and assessed its classification performance. To this purpose, the authors presented a deep learning technique for automated patent categorization based on gated recurrent units and trained word embeddings. Experiments on a conventional evaluation dataset indicated that the strategy improved patent categorization accuracy by 17% compared to state-of-the-art methods.

It is important that the semantic meaning of sentences must be preserved when extracting information[32]. This is the reason why we are adopting the word embedding techniques (word2vec) and Bidirectional Encoder Representations from Transformers (BERT) to preserve the context of sentences during the knowledge extraction from the regulatory text in AEC domain.

It is important to note that in all the aforementioned literature, BERT representations or word embeddings were used as features for their Machine Learning models, assuming that these techniques are capable to represent the semantic regularities within the natural language of regulatory text in AEC domain. However, in our case, the research problem is different – we aim to test the ability of word embeddings and transformers to capture the semantic regularities within the regulatory text in AEC domain, before applying it to downstream reasoning tasks, such as information extraction and rule generalisation. Given that these tasks are critical and the models should be accurate enough to capture the semantics of the language, this paper represents a step ahead what others have done, to make sure that the NLP techniques used to represent the regulatory text are suitable to capture the semantic regularities within the regulatory text in AEC domain.

3. Methodology

This study focuses on modern NLP techniques, namely *word2vec*, *BERT*, and *Sentence BERT* applied to the regulatory text in the construction industry. To this end, we study the capabilities of these techniques to capture the semantic knowledge embedded with the regulatory text at both word and sentence levels. The aim behind this study is to methodically set up the choice of the appropriate NLP techniques as features to represent a domain-specific text, such as the building regulations text. This task is important to guarantee the suitability of these techniques to represent the regulatory text in the AEC domain when they are applied in critical downstream tasks like information extraction or rule generalisation.

3.1. Semantics at word level

In order to capture the semantics at word level from the regulatory text in AEC domain, we propose to use word2vec [1] and BERT [2].

3.1.1. word2vec

Word2vec is a neural network-based approach that uses unsupervised learning to create word embeddings, which are vector representations of words in a high-dimensional space. The algorithm works by training a neural network to predict the context words given a target word or vice versa. In other words, given a large corpus of text data, word2vec learns to represent each word in the corpus as a vector in a multi-dimensional space, such that words that are semantically similar are placed closer together in this space [23].

Word2vec has two main architectures: the *Continuous Bag-of-Words (CBOW)* and the *Skip-gram* model. In CBOW, the algorithm predicts the target word from its context, while in Skip-gram, the algorithm predicts the context words given a target word. Previous results reported in the literature have shown that Skip-Gram [23] model does not only produce useful word representations, but it is also efficient to train. For this reason, we focus on it to build our embeddings for regulatory text in AEC domain in this study.

However, one draw back of the word2vec model is that it does not take into context the use of a word in a sentence as words sometimes have different meanings when applied in sentences. For example, *(i) worker's right in the the site* and *(ii) Right side of the building*. Word2vec will assign similar vectors to the word *"right"*. This gave rise to BERT as discussed in the next section.

3.1.2. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based neural network model for NLP tasks. It was developed by Google AI Language in 2018 and is one of the most popular and powerful pre-trained language models [2]. Unlike traditional NLP models, which process text in a linear manner (from left to right or right to left), BERT is designed to process text in both directions, using a bidirectional approach. This allows BERT to capture the context and meaning of words more accurately, and to better understand the relationships between words and sentences [33].

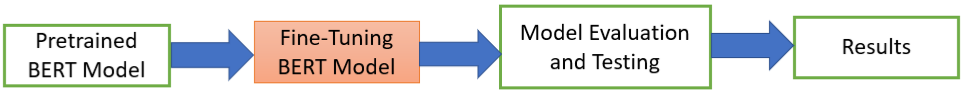


Figure 1: BERT methodology implementation

Figure (1) displays the BERT methodology we used. The BERT model comes pre-trained with about 110Million words[34]. We import the python library *pytorch_transformer*, from which we import a *BertTokenizer* and the pre-trained *BertModel*. Then, we fine-tune the model with our UK building regulations corpora in order to achieve a better performance.

Transformers offers a variety of classes for using BERT on various tasks (token classification, text classification, *etc*). Here, we are utilising the fundamental *BertModel*, which is a decent option if all we want to do with BERT is extract embeddings and has no specified output requirement. We evaluate the model and test how the model predicts missing words in sentences by replacing those words with the [MASK] function. We tested both the pre-trained model and the fine-tuned model in order to test the importance of fine-tuning BERT model in our task.

3.2. Semantics at sentence level

In order to capture the semantics at sentence level from the regulatory text in AEC domain, we propose to use Sentence BERT (SBERT) [24].

3.2.1. Sentence BERT

Sentence-BERT (SBERT) is a variation of the BERT (Bidirectional Encoder Representations from Transformers) model that is specifically designed to generate sentence embeddings, which are vector representations of sentences in a high-dimensional space [24]. The idea behind SBERT is to leverage the power of pre-trained transformer-based models like BERT for sentence-level tasks, such as semantic similarity - the task we are focusing on in this work. Unlike BERT, which generates a fixed-length vector representation for each word in the input sequence, SBERT generates a fixed-length vector representation for each sentence in the input. SBERT achieves this by applying pooling techniques such as max-pooling or mean-pooling over the output of the last layer of the transformer, which results in a sentence-level embedding [35].

Figure (2) displays the process methodology for the SBERT model. First, we converted our PDF to TEXT file and preprocessed the data in the text file using *nltk* to tokenize the sentences.

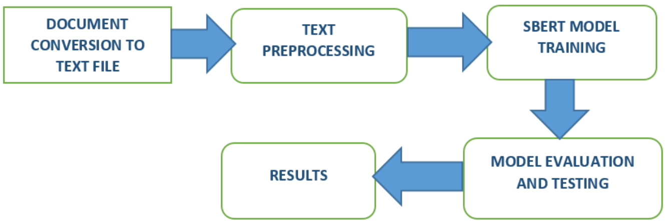

DOCUMENT CONVERSION TO TEXT FILE
TEXT PREPROCESSING
SBERT MODEL TRAINING
MODEL EVALUATION AND TESTING
RESULTS

Figure 2: Sentence BERT methodology implementation

Then, regular expression command was used to clean the data but this time, we did not remove stop words so that we could maintain the semantic relationship between words in the sentence. The output of preprocessing was corpora with only sentences.

The Sentence transformer model SBERT was trained with our dataset to create embeddings. Then we queried the model to extract sentences related to the query, and the results only extracted words and sentence similar to the search query.

3.3. Semantic regularities in the building regulations text

Word embeddings and transformers gain their success from their ability to capture syntactic and semantic regularities in the natural language. Interestingly, they represent each relationship by a relation-specific vector offset [36]. For example, the famous analogy *"king is to queen as man is to woman"* is encoded in the vector space by the vector arithmetic *"king - man + woman = queen"*. More specifically, the word analogy task aims at answering the question *"man is to woman as king is to — ?"* given the two pairs of words that share a relation (*"man:woman"*, *"king:queen"*), where the identity of the fourth word (*"queen"*) is hidden.

Motivated by the ability of modern NLP techniques to extract semantic knowledge in textual data without any prior domain knowledge, this ability is evaluated in a domain-specific text, namely, the regulatory text in the AEC domain. The aim is to assess as to what extent these NLP techniques are able to correctly represent the semantic knowledge in regulatory text given the complexity of the regulations in the construction industry comparing to natural language.

The semantic extraction methodology adopted at the word level is to query for building-related regularities captured in the vector model through simple vector subtraction and addition. More formally, given two pairs of words ($word\_a : word\_a'$) and ($word\_b : word_b'$), the aim is to answer the question *(word_a is to word_a' as word_b is to —?)*. Thus, the vector of the hidden word $word\_b'$ will be the vector ($word\_a' - word\_a + word\_b$), suggesting that the analogy question can be solved by optimising:

$$\arg \max_{word\_b' \in W} (similarity(word\_b', word\_a' - word\_a + word\_b)) \tag{1}$$

where W is the vocabulary and *similarity* is the cosine similarity measure.

This task is challenging for building text language as no gold standard is available to evaluate the efficacy of word embeddings and transformers in identifying linguistic regularities in unstructured regulatory text in AEC domain, unlike existing work that use either the gold standard defined by Mikolov et *al.* [36] for general natural language tasks or predefined ontologies like

NDF-RT ontology[2] for medical domain. Although various building-related ontologies exist, such as ifcOWL [37], the Building Topology Ontology (BOT) [38], the Building Product Ontology (BPO) [39], *etc.*, the automatic mapping between the terminology used in our data sources (UK Regulations) and the ontology concepts was hard and infeasible in most cases. To overcome this problem, we propose to use Brick Ontology [25], which is a semantic metadata standard representing the physical and logical entities in buildings, and a minimal set of relationships that capture the connections between entities. Brick ontology was useful because it essentially replaces the unstructured labels with semi-structured set of tags, which guarantees - to some extent - the mapping between the concepts existing in our corpora and these tags. To build our ground truth, we manually curate relationships related to building regulations domain from Brick Ontology, and define a test set of analogy questions as *semantic questions* following the relation described above, after verifying that the concepts present in every semantic question exist in our corpora. This verification step is necessary to guarantee that all the extracted relationships can be tested and fairly assess the performance of our models. The semantic questions are formed based on the semantic relationships between concepts in the ontology, such as *"is-a"*, *"hasPart"*, *"isPartOf"*, *"isContainedIn"*, *"isTypeOf"*, etc. For example, *"roof"* and *"parapet"* are considered two components of the building elements *"wall"* and *"balcony"*, respectively. Accordingly, the analogical question should be *"roof is part of wall as parapet is part of —?"*. To correctly answer the question, the model should identify the missing term with a correspondence counted as a correct match by finding the word *"balcony"*, whose vector representation is closest to the vector *("roof" - "wall" + "parapet")* according to the cosine similarity. Similarly, for the semantic relationship *"room is a type of space as door is a type of fitting"*, given the terms *"room"*, *"space"*, and *"door"*, the model should be able to predict the term *"fitting"*. Recall that for the specificity and complexity of scientific language and respecting the interchangeability of scientific terms, instead of using the exact correspondence as the correct match, it is proposed to adopt an approximate correspondence that considers an answer as correct if it belongs to the top 10 nearest words given by cosine similarity in order to guarantee the applicability of the generated embeddings in regulatory text in AEC domain. This approach was based on the work published by Dridi et *al.* [22] on word2vec hyper-parametrisation is the scientific domain.

The methodology described above is applied to the word-level semantics. However, for the sentence-level semantics, it is proposed to query the SBERT model with sentences extracted the manual for the UK building regulations, and to test its ability to capture the semantic meaning of sentences, including their context and relationships with other sentences.

4. Experimental Evaluation

4.1. Dataset

To show how word embedding techniques and transformers could extract semantic from text data, we used the UK Building Regulations Code[3]document, which is publicly available and saved in PDF format. The document contains 18 chapters, each of them is related to a specific

[2] National Drug File -Reference Terminology

[3] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082748/Merged_Approved_Documents__Jun2022_.pdf

building domain such as construction, fire safety, energy efficiency, etc. This document contains building rules to guide designers and builders on the things they must adhere to in terms of specifications and guidelines. All chapters in the document were used for our work. The document has been first converted to a text file. Then, it has been pre-processed before training the NLP models. The pre-processing consists of (i) removal of all punctuation and lower-casing the corpus; (ii) removal of stop-words using Stanford NLP stop word list, enriched by a list of irrelevant keywords extracted the UK building regulations code like *"online version, edition, for use in England, etc.*; and (iii) construction of bag-of-words where words are either unigrams used for standard word2vec training or bigrams used for word2phrase learning. The use of bigrams is justified by their frequent use in the building text, such as *"fire safety"*, *"energy efficiency"*, *etc.* Table 1 summarises the statistics of the dataset.

Table 1
UK building regulations code statistics

Document Title	#Chapters	#Sentences	#Words	File Type
HM-Government-The building regulations 2010-The Merged approved document	18	21,337	528,419	PDF

4.2. Results and Discussion

The evaluation of the chosen NLP techniques to extract the semantic knowledge from the building regulations text at both word and sentence levels was performed following the methodology described in Section 3.3.

As described in Section 3.3, the word-level semantic task attempts to query for regularities captured in the embedding model through simple vector subtraction and addition for word2vec and BERT models. The preliminary analogy dataset we created contains 100 analogical questions extracted from the Brick Ontology as described in Section 3.3, and it is considered as our ground truth. The dataset is made publicly available for further use and enrichment [4]. After querying the models with the selected words/phrases, we calculate the similarity scores between the query embeddings and the results embeddings using cosine similarity metric. This gives us a measure of how similar each returned word/phrase from the UK building regulations code is to the query. A summary of our models accuracy in displayed in Table 2. For BERT, both the pre-trained and fine-tuned models were used.

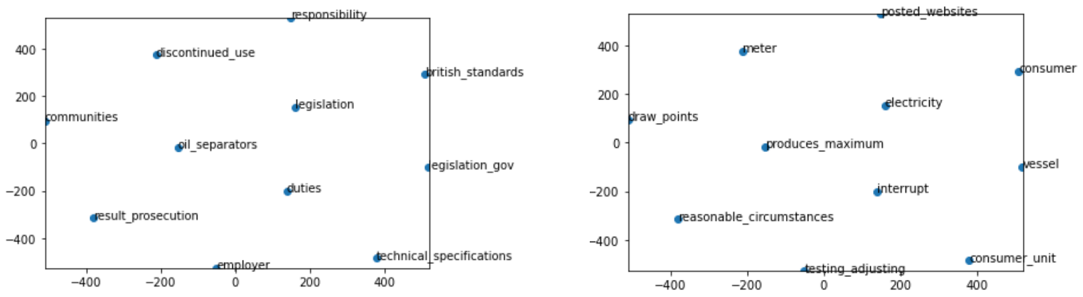
Table 2
Semantic accuracy models at both word and sentence levels

Model	Accuracy(%)
word2vec	61.35%
Pre-trained BERT	54.6%
Fine-tuned BERT	80%
Sentence BERT	100%

[4] https://github.com/salomena/Word-Embeddings-/blob/Semantic-relationships/Semantic%20Relationships

Table 2 shows that the accuracy of word2vec and BERT models are both promising and proves the ability of these two models to capture the semantic regularities within the regulatory text in the AEC domain. Both Word2vec and the fine-tuned BERT perform better than the pre-trained BERT because they have been trained on our dataset. These findings are insightful and promising because they clearly show that both word embeddings (word2vec) and transformers (BERT and SBERT) are able to capture semantic regularities in AEC-related regulatory text, with an accuracy of 80% for BERT at the word-level. Although, it was observed in [40] that BERT performed poorly in domain specific words.

In addition to this quantitative analysis, a qualitative analysis has been performed with word2vec model. It consists to represent the t-distributed stochastic neighbor embedding (t-SNE) [41] visualisations of words.



(a) word "legistlation"

(b) word "electricity"

Figure 3: Vector offsets examples of semantic relationships in the building regulations text

For instance, Figures 3a and 3b represent the vector offsets of the two words *"legistlation"* and *"electricity"*, respectively. It can be clearly seen from the plots that the surrounding words of each word are semantically closer in meaning. This confirms that modern NLP techniques, namely word2vec and BERT can reliably extract the semantic knowledge from the building regulations text.

For the sentence-level semantic task, SBERT has been evaluated as described in Section 3.3. A set of sentences has been selected from the manual for the UK building regulations as queries, the semantic search has been performed, and then the returned results have been evaluated with cosine similarity measure. Interestingly, the model hits a 100% accuracy at the sentence-level as shown in Table 2. These results are very promising and confirm the previous findings on the capability of word embeddings and transformers to capture the semantic knowledge at both word and sentence levels.

5. Conclusions and Future Work

Despite being popular and achieving state-of-the-art performance in tasks related to knowledge extraction and semantic search, word embeddings and transformers are still being used intuitively; without a proper testing of their ability to capture semantic regularities in a domain-specific text. From this perspective and aiming to provide a reliable information extraction from the regulatory text in AEC domain, this work explored three models, namely word2vec, BERT and Sentence BERT and tested their reliability for the extraction of semantics from the building text at both word and sentence levels. The UK building regulations code has been used as a dataset to apply and test the models, and Brick Ontology has been used as a ground truth the create the semantic relationships. The obtained results were insightful and promising. This work adds breadth to the automation in construction industry that started to heavily rely on ML and NLP techniques to deal with the massive amount of textual data. Applied to the regulatory text, and despite the sensitivity of the domain, our work has proven the ability of the modern NLP techniques to effectively capture the semantic knowledge.

As a short term objective, we plan to expend our ground truth of semantic analogies in the building domain by combining the knowledge extracted from different resources, and see how the models perform on a larger dataset. For long term objectives, we plan to leverage these NLP techniques to extract information and auto-generate rules from the building regulations.

6. Acknowledgements

This work is partially funded by the European Union's Horizon Europe research and innovation programme under grant agreement no 101056973 (ACCORD). UK Participants in Horizon Europe Project [ACCORD] are supported by UKRI grant numbers [10040207] (Cardiff University), [10038999] (Birmingham City University and [10049977] (Building Smart International).

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, 2013, pp. 3111–3119.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[3] S. Shrestha, S. Morshed, N. Pradhananga, X. Lv, Leveraging accident investigation reports as leading indicators of construction safety using text classification, in: Conference: ASCE Construction Research Congress (CRC) 2020, 2020, pp. 490–498.

[4] D. Salama, N. El-Gohary, Semantic modeling for automated compliance checking, in: Computing in Civil Engineering (2011), 2011, pp. 641–648.

[5] J. Zhang, N. El-Gohary, Extraction of construction regulatory requirements from textual documents using natural language processing techniques, in: Computing in Civil Engineering (2012), 2012, pp. 453–460.

[6] J. Zhang, N. El-Gohary, Information transformation and automated reasoning for automated compliance checking in construction, Computing in civil engineering 8 (2013) 701–708.

[7] J. Zhang, Automated code compliance checking in the construction domain using semantic natural language processing and logic-based reasoning, University of Illinois at Urbana-Champaign, 2015.

[8] J. Zhang, N. M. El-Gohary, Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking, Journal of Computing in Civil Engineering 30 (2016) 04015014.

[9] Y. Zhang, R. He, Z. Liu, K. H. Lim, L. Bing, An unsupervised sentence embedding method by mutual information maximization, arXiv preprint arXiv:2009.12061 (2020).

[10] L. Zhang, J. Wang, Y. Wang, H. Sun, X. Zhao, Automatic construction site hazard identification integrating construction scene graphs with bert based domain knowledge, Automation in Construction 142 (2022) 104535.

[11] D. Boyd, Social media: A phenomenon to be analyzed, Social Media+ Society 1 (2015) 2056305115580148.

[12] L. Shen, H. Yan, H. Fan, Y. Wu, Y. Zhang, An integrated system of text mining technique and case-based reasoning (tm-cbr) for supporting green building design, Building and Environment 124 (2017) 388–401.

[13] N. Jung, G. Lee, Automated classification of building information modeling (bim) case studies by bim use based on natural language processing (nlp) and unsupervised learning, Advanced Engineering Informatics 41 (2019) 100917.

[14] J. Padhy, M. Jagannathan, V. Delhi, Application of natural language processing to automatically identify exculpatory clauses in construction contracts, Journal of Legal Affairs and Dispute Resolution in Engineering and Construction 13 (2021) 04521035.

[15] S. M. Weiss, N. Indurkhya, T. Zhang, F. Damerau, Text mining: predictive methods for analyzing unstructured information, Springer Science & Business Media, 2010.

[16] S.-H. Hong, S.-K. Lee, J.-H. Yu, Automated management of green building material information using web crawling and ontology, Automation in Construction 102 (2019) 230–244.

[17] Y. Jallan, E. Brogan, B. Ashuri, C. Clevenger, Application of natural language processing and text mining to identify patterns in construction-defect litigation cases, Journal of Legal Affairs and Dispute Resolution in Engineering and Construction 11 (2019) 04519024.

[18] F. C. Pereira, F. Rodrigues, M. Ben-Akiva, Text analysis in incident duration prediction, Transportation Research Part C: Emerging Technologies 37 (2013) 177–192.

[19] F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys (CSUR) 34 (2002) 1–47.

[20] M. Al Qady, A. Kandil, Automatic clustering of construction project documents based on textual similarity, Automation in Construction 42 (2014) 36–49.

[21] F. Hutter, H. Hoos, K. Leyton-Brown, An efficient approach for assessing hyperparameter importance, in: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14, JMLR.org, 2014, p. I–754–I–762.

[22] A. Dridi, M. M. Gaber, R. Azad, J. Bhogal, k-nn embedding stability for word2vec hyperparametrisation in scientific text, in: International Conference on Discovery Science, Springer, 2018, pp. 328–343.

[23] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[24] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[25] G. Fierro, J. Koh, Y. Agarwal, R. K. Gupta, D. E. Culler, Beyond a house of sticks: Formal-

izing metadata tags with brick, in: The 6th ACM International Conference on Systems for EnergyEfficient Buildings, Cities, and Transportation, Association for Computing Machinery, New York, NY, USA, 2019, p. 125–134.

[26] R. Zhang, N. El-Gohary, A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking, Automation in Construction 132 (2021) 103834.

[27] J. A. Miñarro-Giménez, O. Marín-Alonso, M. Samwald, Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation, CoRR abs/1502.03682 (2015). arXiv:1502.03682.

[28] Y.-C. Zhou, Z. Zheng, J.-R. Lin, X.-Z. Lu, Integrating nlp and context-free grammar for complex rule interpretation towards automated compliance checking, Computers in Industry 142 (2022) 103746.

[29] A. J.-P. Tixier, M. Vazirgiannis, M. R. Hallowell, Word embeddings for the construction domain, arXiv preprint arXiv:1610.09333 (2016).

[30] W. Yuan, Y. Lei, X. Guo, Research on text similarity calculation based on bert and word2vec, in: ICETIS 2022; 7th International Conference on Electronic Technology and Information Science, VDE, 2022, pp. 1–4.

[31] J. Risch, R. Krestel, Domain-specific word embeddings for patent classification, Data Technologies and Applications (2019).

[32] D. Tian, M. Li, Q. Ren, X. Zhang, S. Han, Y. Shen, Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining, Automation in Construction 145 (2023) 104670.

[33] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[35] B. Wang, C.-C. J. Kuo, Sbert-wk: A sentence embedding method by dissecting bert-based word models, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2146–2157.

[36] T. Mikolov, Y. Wen-Tau, Z. Geoffrey, Linguistic regularities in continuous space word representations., in: HLT-NAACL, 2013, pp. 746–751.

[37] P. Pauwels, W. Terkaj, Express to owl for construction industry: Towards a recommendable and usable ifcowl ontology, Automation in Construction 63 (2016) 100–133.

[38] M. H. Rasmussen, M. Lefrançois, G. Schneider, P. Pauwels, Bot: the building topology ontology of the w3c linked building data group, Semantic Web (2020) 143–161.

[39] A. Wagner, W. Sprenger, C. Maurer, T. E. Kuhn, U. Rüppel, Building product ontology: Core ontology for linked building product data, Automation in Construction 133 (2022) 103927.

[40] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).

[41] L. van der Maaten, G. E. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.