LEVERAGING WORD EMBEDDINGS AND TRANSFORMERS TO EXTRACT SEMANTICS FROM BUILDING REGULATIONS TEXT

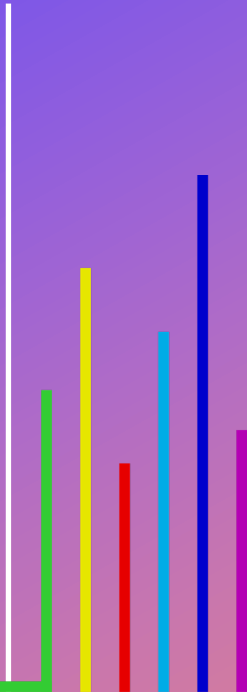Presenter: Edlira Vakaj, Edlira.Vakaj@bcu.ac.uk

Authors: Odinakachukwu Okonkwo, Amna Dridi and Edlira Vakaj

Birmingham City University, School of Computing and Digital Technology

LDAC2023 - Linked Data in Architecture and Construction Week

Aim

- Study the performance of modern NLP techniques on extracting semantic regularities in AEC

- Word embeddings and transformers
  - word2vec
  - BERT
  - Sentence BERT (SBERT)

- The UK building regulations code has been used as a dataset
- BRICK Ontology
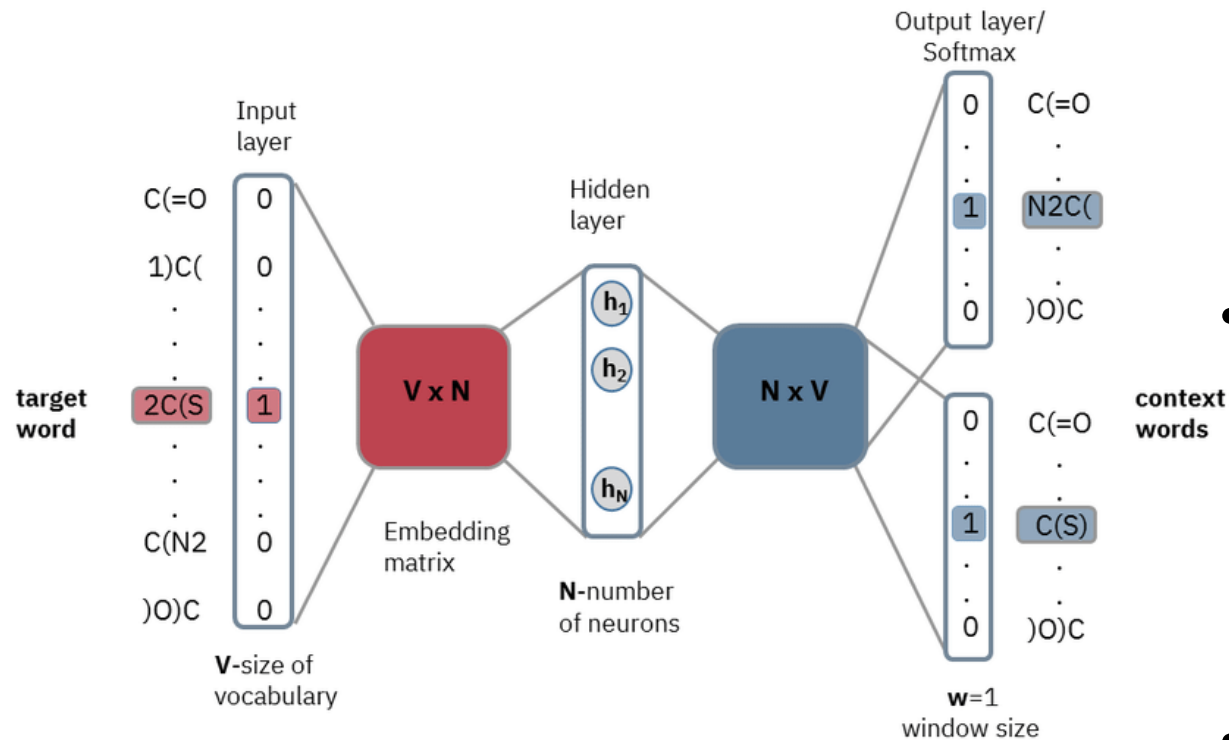
Semantic Regularities in NLP

What?

- Semantic Regularities: underlying structure or patterns that exist in the meaning or semantics of language

- NLP tasks examples: word sense disambiguation, semantic role labeling, and semantic similarity estimation

Why?

- Information Retrieval

Semantic relatedness allows for a deeper understanding of the user's intent, helping to match their query with the most relevant content.

word2vec

- Neural network-based approach to create word embeddings

- Two main architectures
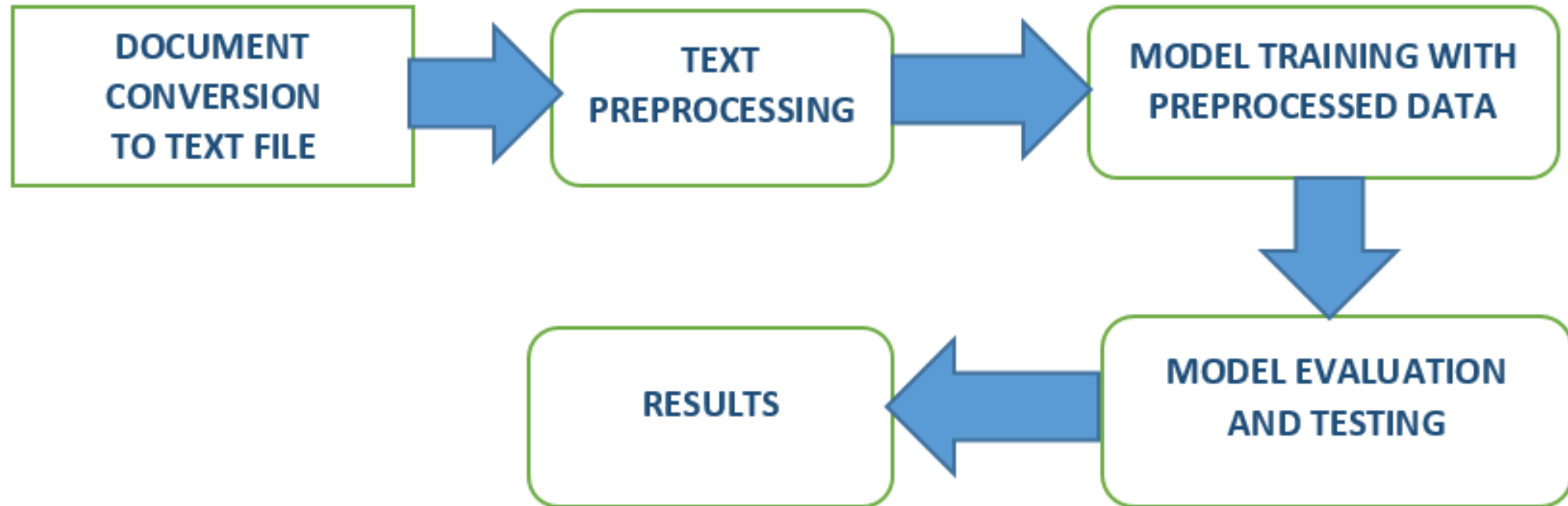 - Continuous Bag-of-Words (CBOW)
  - Skip-Gram

- Lack of context

BERT

- BERT(Bidirectional Encoder Representation from Transformers)
- Process text in both directions allowing capture of context
- Pretrained (110 million words)
- Fine tuned

BERT methodology implementation

Sentence BERT (SBERT)

- SBERT a variation of BERT for sentences

Sentence BERT methodology implementation

Semantic regularities in the building regulations text

- Relation specific vector: (*word_a : word_a'*) and (*word_b : word_b'*)
    - *"kind is to queen as man is to <u>woman</u>"*
    - *"room is a type of space as door is a type of <u>fitting</u>"*

$$\text{arg max }_{word\_b' \in W}(similarity(word\_b', word\_a' - word\_a + word\_b))$$

Data Set

- BRICK Ontology as ground truth
- 100 sentences

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

GitHub Link: https://github.com/salomena/Word-Embeddings-/tree/Semantic-relationships

GitHub Resources

NLP Pipelines


```
Word-Embeddings- / Project Code

Code    Blame    398 lines (282 loc) · 11.1 KB

38    corpus = nltk.sent_tokenize(contents)
39    #print(corpus)
40
41    #Preprocessing Commenced using regular expression
42    for i in range(len(corpus)):
43        corpus [i] = corpus [i].lower()    #converts all Uppercase letters to small letters
44        corpus [i] = re.sub(r'\W',' ',corpus [i])  #removed empty spaces
45        corpus [i] = re.sub(r'\s+',' ',corpus [i]) #removed punctuation marks
46        corpus [i] = re.sub(r'[^A-Za-z]',' ',corpus [i]) #removed special characters like @
47        shortword = re.compile(r'\W*\b\w{1}\b') # remove short words
48        corpus [i] = re.sub(shortword,' ',corpus [i])
49    #print(corpus)
50
51
52    #removing stop words from corpus and saving output text to file
53    from nltk.corpus import stopwords
54    stopwords.words('english')
55    stop_words = set(stopwords.words('english'))
56    file1=open("HMbuildingclean.text","a",encoding='utf-8')
57    for sentence in corpus:
58        tokens = nltk.word_tokenize(sentence)  #Tokenization
59        for token in tokens:
60            token=''.join([i for i in token if not i.isdigit()])
61            if not token in stop_words:
62                file1.write(" "+token)
```
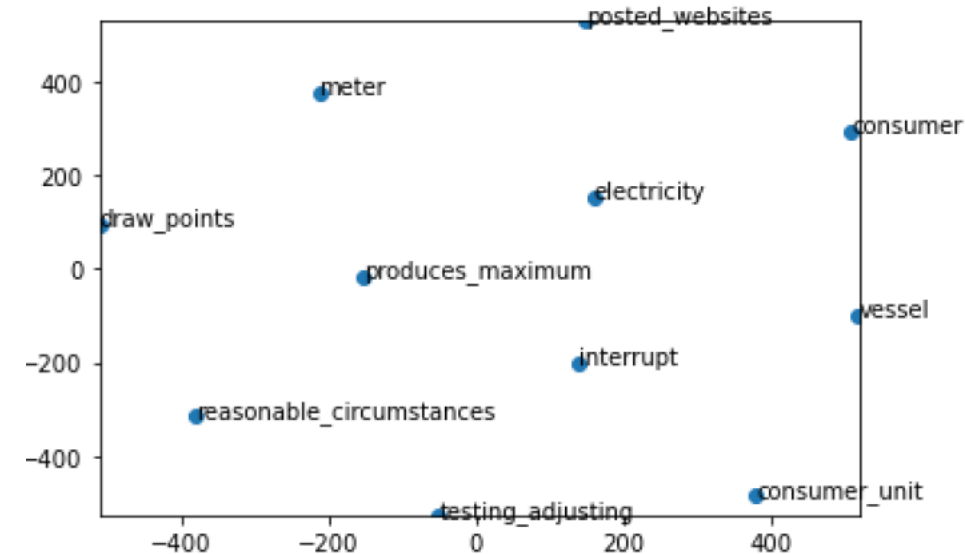
Brick enabled ground truth

A Photovoltaic array collects Solar Energy as roof collects rainwater.
A construction of Photovoltaic panels is called a photovoltaic array as a con
A heat pump can form an air loop as a water pump can form a water loop.
Exhaust outlet is part of the ventilation system as Hot water pipe is part of
The boiler is part of a domestic hot water system as flow sensor control is p
Hot water pipe is part of a domestic hot water system as flow sensor control
Solar Energy is stored in a battery as as waste water is stored in a septic t
Chilling Plant provides chilled water as the boiler provides hot water.
Outdoor fans is contained in cooling towers as switches is contained in a cir
Heat pump is used to capture heat as photovoltaic panels capture solar energy
A switch controls the lighting system as daylight control regulates natural l
daylight control regulates natural lighting as Ventilators are used to regula
the amount of light falling on a surface is termed illuminance as the amount
The emergency assistance alarm is activated by a pull cord as Pull cords are
Automatic fire detection triggers the fire alarm as smoke dectectors trigger
Electricty use is measured by the electric meter as gas use is measured by a
Pull cords are used to control blinds as switches as used to control lighting

Results and conclusions

Quantitative

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

Qualitative (word2vec)

t-SNE visualisation of the word "electricity"

Results and conclusions


```
In [35]: query = ('thermally activated fire dampers')
    ...:
    ...: queries=[query]
    ...:
    ...: query_embeddings = model.encode(queries)
    ...:
    ...: print("Semantic search result")
    ...:
    ...: for query, query_embedding in zip(queries, query_embeddings):
    ...:     distances = scipy.spatial.distance.cdist([query_embedding], sentence_embeddings_base, "cosine")[0]
    ...:
    ...:     results = zip(range(len(distances)), distances)
    ...:     results = sorted(results,key=lambda x: x[1])
    ...:     print(" ")
    ...:     print("Query:", query)
    ...:     print("\n Top 5 most similar sentences in corpus")
    ...:     print(" ")
    ...:     for idx, distance in results[:5]:
    ...:         print(sentences[idx].strip(), "(Cosine Score: %.4f)"%(1-distance))
Semantic search result

Query: thermally activated fire dampers

 Top 5 most similar sentences in corpus

Method 1 — thermally activated fire dampers. (Cosine Score: 0.9091)
Method 1 — thermally activated fire dampers. (Cosine Score: 0.9091)
Fire extinguishers. (Cosine Score: 0.8548)
Fire extinguishers. (Cosine Score: 0.8548)
Flueblock chimney systems consist of a set
from the thermal radiation from a fire. (Cosine Score: 0.8442)

In [36]: query = ('door frames and doors')
```

Sentence BERT

Future work

Regulations

IfcElement

embedding

Building work

Regulation 3 of the Building Regulations defines 'building work'. Building work includes:

a. the erection or extension of a building

b. the provision or extension of a controlled service or fitting

c. the material alteration of a building or a controlled service or fitting.

Regulation 4 states that building work should be carried out in such a way that, when work is complete:

a. for new buildings or work on a building that complied with the applicable requirements of the Building Regulations: the building complies with the applicable requirements of the Building Regulations.

b. for work on an existing building that did not comply with the applicable requirements of the Building Regulations:

   (i) the work itself must comply with the applicable requirements of the Building Regulations

   (ii) the building must be no more unsatisfactory in relation to the requirements than before the work was carried out.

Material change of use

Regulation 5 defines a 'material change of use' in which a building or part of a building that was previously used for one purpose will be used for another.

The Building Regulations set out requirements that must be met before a building can be used for a new purpose. To meet the requirements, the building may need to be upgraded in some way.

Energy efficiency requirements

Part 6 of the Building Regulations imposes additional specific requirements for energy efficiency.

If a building is extended or renovated, the energy efficiency of the existing building or part of it may need to be upgraded.

THANK YOU

QUESTIONS ARE
WELCOMED