SPAR.txt, a cheap Shallow Parsing approach for Regulatory texts

Ruben Kruiper[*] and **Ioannis Konstas** and **Alasdair Gray**
School of Mathematics and Computer Sciences
Heriot-Watt University, Edinburgh, United Kingdom
*ruben.kruiper@hw.ac.uk

Farhad Sadeghineko and **Richard Watson** and **Bimal Kumar**
Department of Architecture and Built Environment
Northumbria University, Newcastle, United Kingdom

Abstract

Automated Compliance Checking (ACC) systems aim to semantically parse building regulations to a set of rules. However, semantic parsing is known to be hard and requires large amounts of training data. The complexity of creating such training data has led to research that focuses on small sub-tasks, such as shallow parsing or the extraction of a limited subset of rules. This study introduces a shallow parsing task for which training data is relatively cheap to create, with the aim of learning a lexicon for ACC. We annotate a small domain-specific dataset of 200 sentences, SPAR.txt[1], and train a sequence tagger that achieves 79,93 F1-score on the test set. We then show through manual evaluation that the model identifies most (89,84%) defined terms in a set of building regulation documents, and that both contiguous and discontiguous Multi-Word Expressions (MWE) are discovered with reasonable accuracy (70,3%).

1 Introduction

Non-compliance with building regulations has been linked to fatal incidents (Cook, 2017). However, ensuring that a building complies with regulations is complicated and time-consuming because:

- Regulations contain ambiguous and sometimes conflicting criteria (Cook, 2017; Hywel et al., 2020).

- Regulations change and are distributed over many documents (Fuchs, 2021), e.g., over 800 documents in the U.K. with many of them behind a paywall.

- Criteria often refer to entire sections in other documents, e.g., "*The emergency lighting should be installed in accordance with BS 5266: Part 1: 2016 as read in association*

with BS 5266: Part 7: 1999 (BS EN: 1838: 2013)." (Scottish Government, 2020).

- Regulations differ per country, and some criteria borrowed from international regulations are not suited to the specific environment (Moon et al., 2019).

Automated Compliance Checking (ACC) could reduce the difficulty, time, costs and number of human errors made during compliance checking (Dimyadi and Amor, 2013; Preidel and Borrmann, 2018), as well as ease customisation and innovation in the building sector (Niemeijer et al., 2014). There exist two branches of ACC research. One focuses on tools that reason over a rule-base – often consisting of hard-coded rules (Pauwels et al., 2017; Solihin et al., 2019). The other branch attempts semantically parse the Natural Language regulations into rules that enable reasoning – a complex task that is of interest to the wider legal domain (Wyner et al., 2012).

This study presents a novel *shallow parsing* task, for which the creation of training data is cheap, and an accompanying small dataset of 200 sentences. The aim is to learn a semantic lexicon for ACC, which is often an important first step for semantic parsing because it enables the grounding of information units identified in a text – such as objects, interactions and constraints (Zettlemoyer and Michael, 2005; Kollar et al., 2010; Chen, 2012). Section 2 motivates our task and describes related work with a focus on parsing building regulations. 3 describes the task, as well as the collection of a small annotated dataset – SPAR.txt – for the task of discovering and identifying domain-specific terms, including Multi-Word Expressionss (MWE). In 4 we describe and train a sequence tagging model, which generalises well to unseen text within the same domain. 5 describes the evaluation of outputs, specifically with regards to the objects identified in a corpus of 13K sentences derived from the Scot-

---

[1]For the SCOTREG corpus, SPAR.txt dataset and code see: https://github.com/rubenkruiper/SPaR.txt

tish Building Regulations (Scottish Government, 2020). Our model finds 89,84% of the terms explicitly defined in these documents. Furthermore, a significant proportion (70,3%) of these predictions matches exactly what a human annotator would consider to be an object in a given sentence context. We argue that our new task can provide a cheap approach to lexicon learning that could benefit (1) Information Extraction (IE) in support of ACC, (2) Information Retrieval (IR) in support of manual compliance checking, and (3) the mapping of unstructured text to a structured representation.

2 Related work

2.1 Semantic parsing for ACC

Semantic parsing revolves around learning the meaning of Natural Language and converting it to an executable logical form, which is a hard and unsolved task (Mooney, 2007; Artzi and Zettlemoyer, 2013). The fragmented structure of legal texts further complicates semantic parsing in legal domains (Lawsky, 2017), e.g., a clause may state that some object must comply with all of the regulations in some other section or document. Therefore, statutory reasoning requires defeasible logic representations that allow conclusions to be defeated on the basis of subsequent information (Pertierra et al., 2017). Due to the complexity of the task, existing approaches to semantic parsing in the ACC domain often limit their scope to parsing quantitative requirements in some small sub-domain (Fuchs, 2021; Moon et al., 2021).

In the ACC domain, studies on semantic parsing rely on traditional Rule-Based Systems to extract concepts and construct logical statements, e.g., (Zhang and El-Gohary, 2016b,a; Zhou and El-Gohary, 2019; Xu and Cai, 2019). Empirical approaches may be able to handle the combinatorial explosion caused by ambiguity in Natural Language (Wyner et al., 2012). But training data for such systems is difficult and costly to collect (Chen, 2012; Herzig and Berant, 2017). Unsurprisingly, Machine Learning studies in the ACC domain focus on simpler sub-tasks of semantic parsing for which training data is relatively easy to collect, such as Named Entity Recognition (Liu and El-Gohary, 2017; Moon et al., 2021) and Relation Classification (Zhong et al., 2020). This study follows a similar strategy to limit the complexity of annotation, see 3.2 and 3.3.

2.2 Shallow parsing for ACC

Orthogonal research on shallow parsing in the ACC domain includes (1) the decomposition of complex sentences into parts that are easier to process (Zhang and El-Gohary, 2019), and (2) more-or-less idiosyncratic semantic markup schemes that help identify requirements and their components in text, e.g., (Hjelseth and Nisbet, 2011) and (Zhang and El-Gohary, 2016b). Efforts to automate such shallow parsing approaches encountered performance issues when handling MWEs (Zhang and El-Gohary, 2019; Zhang and Nora, 2020).

Although the proper handling of MWEs is a key issue in Natural Language Processing (Siskind, 1996; Sag et al., 2002; Ramisch et al., 2018), extracting MWEs is especially relevant to IE in domains rich in technical terms (Baldwin and Kim, 2010) – such as the building regulations. Processing MWEs is a general requirement for ACC, because both single and multi-word concepts mentioned in regulations have to be aligned with components and values found in Building Information Model (BIM) models. Such BIM models rely on standards, such as the Industry Foundation Classes (IFC) data model, to facilitate amongst others compliance checking and information exchange between applications and potentially international stake-holders (Plume and Mitchell, 2007; Beetz et al., 2009; Pauwels et al., 2017). This study aims to automatically learn a vocabulary for ACC, which entails MWE processing.

2.3 Processing MWEs

Processing MWE can be divided into two sub-tasks: MWE discovery and MWE identification (Constant et al., 2017).

MWE discovery aims to find new types of MWEs in text corpora and storing them in a lexicon. Unsupervised approaches are used that rely on properties of MWEs that set them apart from random combinations of words, such as word collocation frequency (Manning et al., 2002; Pecina and Schlesinger, 2006), non-substitutability of component words (Lapata and Lascarides, 2003; Constant et al., 2017), and non-compositionality (Frege, 1996; Riedl and Biemann, 2015). The latter applies mostly to idiomatic expressions(Villavicencio and Idiart, 2019), such as '*cloud nine*'.

MWE identification revolves around annotating MWEs in a corpus, based on a lexicon or on results from MWE discovery (Constant et al., 2017).

	Domestic	Non-Domestic	Total
Terms defined in definitions section	128	127	128
Defined terms in text after lemmatisation and lower-casing	233	247	292
Number of terms linking to definitions section	4.687	5.368	10.055
Number of tokens	131.666	151.499	283.165
Vocabulary	8.285	8.925	9.837
Number of sentences	6.313	7.293	13.606
Mean sentence (word-level token) length, excluding punctuation	20,86	20,77	20,81
Standard deviation	11,96	12,32	12,16

Table 1: Statistics for the SCOTREG corpus – the number of defined terms, word-level tokens and sentences found in the domestic and non-domestic Scottish Building regulations.

This enables representing MWEs as single tokens, which has been shown to improve accuracy of NLP tasks (Green et al., 2011), such as dependency parsing (Nivre and Nilsson, 2004). Supervised approaches are used, amongst which sequence tagging has been found to work well (Constant et al., 2017), e.g., (Blunsom and Baldwin, 2006; Constant et al., 2012). Sequence tagging has also been used for joint MWE identification and Part-of-Speech (POS) tagging (Constant and Sigogne, 2011) and may be amended to handle discontiguous MWEs (Schneider et al., 2014).

This study explores sequence tagging for joint MWE processing. However, this study does not aim to handle idiomatic expressions or proverbs. Beyond research on MWE processing, a related task that focuses on identifying technical terms and Named Entities is concept mining, e.g., (Rajagopal et al., 2013; Poria et al., 2014). In contrast to these concept mining studies, we do not rely on dependency parses or external resources, such as ConceptNet (Speer et al., 2013).

3 MWE tagging for ACC

3.1 A building regulations corpus

Many building regulations are captured in formats that are not easily processed by computers, such as PDF (Fuchs, 2021). The Scottish Building Regulations (Scottish Government, 2020) are an exception and are openly available online. We scrape the domestic and non-domestic regulations, including text found in lists, side-notes, tables and captions.

We use TextBlob [2] for word-level tokenization (Penn Treebank Tokenizer) and sentence splitting (PunktSentenceTokenizer (Kiss and Strunk, 2006)). We will refer to the resulting corpus as SCOTREG. An overview of the size of SCOTREG in terms of sentences and tokens can be found in Table 1, as well as the number the terms defined in the '*Terms and definitions*' sections. Throughout SCOTREG defined terms provide a hyperlink to the definitions section. This enables us to count how often defined terms occur. Note these defined terms represent *classes*, they are expressed in various ways throughout the texts and some of the variations do not match any of the terms verbatim – even after lower-casing and lemmatizing.

3.2 Problem statement

The IFC schema does not comprehensively cover the terminology used in the building regulations. In "*A roof covering or roof light which forms part of an internal ceiling lining should [...]*"[3] the term '*roof light*' would fall under the more generic IFC class '*window*'. The problem is compounded as building regulations cover a wide range topics (and thus concepts) – from design and construction, including fire regulations and accessibility, to facility management, renovation and demolition (Pauwels et al., 2017). Despite the more fine-grained terminology used in the SCOTREG corpus, only 128 terms are defined – neither '*roof light*', '*roof covering*' nor '*internal ceiling lining*' are defined.

3.3 Task description

The task in this study is to identify low-level constituent parts of a sentence, which we will refer to as *spans*. We assume that contextual interdependencies between single-word and multi-word spans help tackle lexical sparsity of MWEs. Therefore, we tag both single words and groups of words in a sentence and we tag all tokens exhaustively, including punctuation. Whether spans comprise a single punctuation mark or a group of possibly discontiguous tokens, they should represent a coherent unit of grammatical meaning.

---

[3]For an overview of used examples, the full sentences, and origin of these examples see Table 7 in Appendix B.
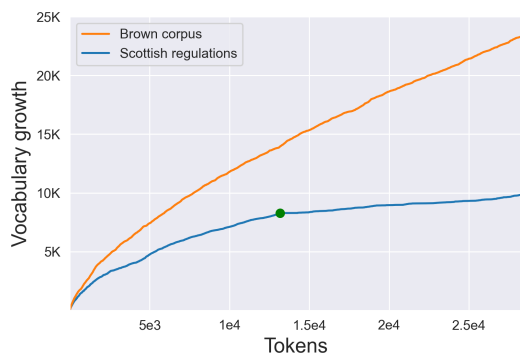
Figure 1: Vocabulary growth for the SCOTREG corpus in comparison to the heterogeneous BROWN corpus. The dot indicates the division between domestic and non-domestic building regulations.

- "*The Building (Scotland) Act 2003 gives Scottish Ministers* [...]"

'*The Building (Scotland) Act 2003*' is a Named Entity that should be treated as a single span. But a regular noun-chunking approach would break on the parentheses and the cardinal number – it may even treat every capitalised word as a proper noun.

- "*A roof covering or roof light which forms part of an internal ceiling lining should* [...]"

A constituency parser may break the term '*roof covering*' into separate words, because a POS tagger would typically assign '*covering*' the tag VBG.

- "[...], *a paved (or equivalent) footpath at least 900mm wide* [...]"

In this case, while we would like to split '*a paved (or equivalent) footpath*' into the spans: '*a paved footpath*', '*(*', '*)*', '*or*' and '*equivalent*'. In a downstream task this would allow us to define a class for '*paved footpath*', and reason over the equivalent types of footpaths. However, to extract '*a paved footpath*' from the sentence above, a discontiguous span representation is required.

3.4 Simplifying assumptions

For our task we rely on a simplified definition of MWEs: "*possibly discontiguous combinations of at least two tokens, where tokens are separated by white-spaces or punctuation in text*" – similar to (Villavicencio and Idiart, 2019). We also make two assumptions on the types of MWEs that we expect to find in the building regulation texts. First, we assume that the building regulations contain few

to no idiomatic expressions, because these may introduce ambiguity. We justify this assumption as the building regulations use a slightly more formal syntax, albeit not as strict as other types of legal text (Chalkidis et al., 2020).

Second, we expect a relatively low variability in surface forms. Both for verbal expressions found in clauses, e.g., '*X should conform to Y*', and for the surface forms of MWEs that indicate technical terms, e.g., '*insulation envelope*' and '*self-contained emergency luminaries*'. A low variability in surface forms would be reflected by a relatively small vocabulary – which is thought to ease the complexity of various NLP tasks (Church, 2013). As can be seen in Figure 1, SCOTREG has in the order of 10K unique tokens for a total of 283K. In comparison to the more heterogeneous Brown corpus (Francis and Kucera, 1964), which has more than 23K unique tokens for the first 283K tokens, SCOTREG indeed has a small vocabulary.

3.5 Annotating SPAR.txt

A domain expert annotated a random selection of 200 sentences in BRAT (Stenetorp et al., 2012). Our assumption is that such a small dataset should suffice for achieving reasonable results on the proposed parsing task. Figure 2 exemplifies how annotations can span single words, multiple words and also indicate that two groups of words belong to a single, discontiguous span. To distinguish between verb-based and noun-based spans, as well as spans that belong to neither of these classes, we annotate the following span types:

- OBJECT spans indicate either real-world objects or distinguishable concepts. They include proper nouns, compounds, multi-word terms, and multi-word Named Entities, such as '*the Target Emissions Rating*', '*offensive fire-fighting*' and '*BS 8000-15: 1990*'. We include determiners as part of the OBJECT span during annotation, see Figure 2.

- ACTION spans may help identify whether a sentence expresses a requirement, similar to (Hjelseth and Nisbet, 2011). We include verbs, support verbs, prepositional verbs and verb-particle constructions, e.g., '*should be maintained*' and '*takes account of*', but we expect to split light-verb constructions, such as '*to take a shower*' (Constant et al., 2017) into '*to take*' and the OBJECT span '*a shower*'.
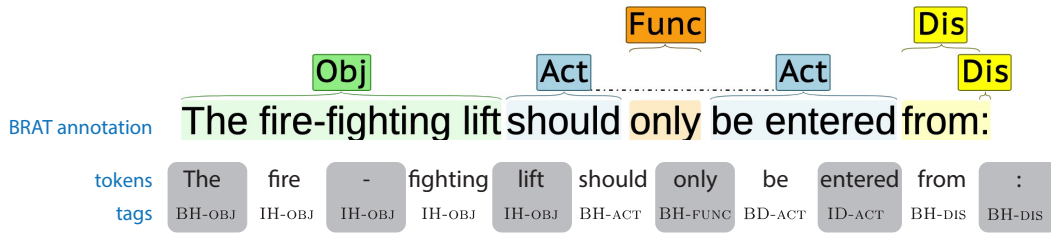
Figure 2: Example of an annotated sentence. The determiner at the start of the OBJECT span is taken to be part of the span. A discontiguous ACTION span is interjected by a FUNCTIONAL span that modifies the Verb-Phrase. During training the sentence is tokenized and the aim is to predict the correct tags for each token, see the tagging scheme described in Section 4.1. The identifier for this sentence in the dataset is 'd_2.14.4_i3_s_0'.

- FUNCTIONAL spans are modifiers that are not inherently part of an OBJECT or ACTION span. They include adverbs and adjectives, e.g., 'main' and 'principal' in 'the main or principal bedroom', as well as complex function words, e.g., 'up to'.

- DISCOURSE spans include punctuation, co-reference anaphora, conjunctions and disjunctions, e.g., ',', 'or' and 'this'.

MWEs can blur the lines between syntax and semantics (Green et al., 2011), with ambiguous cases leading to annotation inconsistencies (Hollenstein et al., 2016). Because the task involves annotating both the words that make up an MWE and those surrounding it, the annotator is forced to come up with more-or-less consistent decisions. As a tool to determine which words belong together, annotators are asked to rely on standard constituency tests:

- Substitution test – if you can substitute a part of a sentence with another word or group of words that belong to the same type, the part is a constituent.

- Pronoun test – if you can replace a part of a sentence with 'it', the part is a constituent.

- Question by repetition test – if you can repeat a part of a sentence, within a valid question, then it is a constituent.

Nevertheless, annotating MWEs often requires domain-specific knowledge and remains ambiguous. Using Figure 2 as an example, one might argue that the preposition 'from' could be part of the AC-TION span 'should be entered' – considering that 'from' converts 'to enter' from an intransitive to a transitive verb. We define several loose guidelines to warrant further consistency in annotation:

1. **Punctuation** Unless punctuation should be part of an OBJECT span, such as the colon in the document name 'BS 800-15:1990', then punctuation should be marked as DISCOURSE.

2. **Negation** Wherever negation is separable from other spans, e.g., '[...] is not level [...]', it should be annotated separately as a FUNC-TIONAL span. In cases where negation is not separable, e.g., in '[...] cannot gain [...]', the current approach is to tag the word containing the negation as FUNCTIONAL.

3. **Granularity** Annotators should rely on domain knowledge to determine whether a Noun-Phrase should be split, e.g., '[...] the effect of smoke travelling along a ceiling [...]' should be broken up while "[...] fire and rescue service personnel [...]' should form a single span.

4. **Coordination** Coordinating conjunctions and disjunctions should be split. Coordinated MWEs may share a span, e.g., Figure 3 shows how a list of 'Standards' share a word that is crucial to the semantics of the individual items. Similarly, conjunctions and disjunctions sometimes share a determiner, e.g., 'the size and orientation of the windows' would include the spans 'the size' and 'the orientation'. The motivation is to help downstream tasks determine that both 'size' and 'orientation', here, are properties of 'the window'.

5. **Overlap** We currently limit the annotation of discontiguous spans to a maximum of two parts. Therefore, overlapping spans should only occur this cannot be avoided, e.g., 'the Silver level' and 'the Gold level' in 'the Silver and Gold level' would need to exist of three single-word spans to avoid overlap.

6. **Adverbs** typically express a manner, place, time, frequency, degree, and so on. The expectation is that such expressions will usually be labelled as FUNCTIONAL, e.g., '*only*' in Figure 2.

7. **Adjectives** typically modify nouns and may or may not be part of an OBJECT span. This decision would be based on whether the modified noun is likely to constitute a separate category or not. As an example, '*structural*' would be part of the OBJECT span '*the structural properties*' but a separate FUNCTIONAL span in '*matters of structural concern*'.

8. **Quantities and units** are treated as a single OBJECT span. If such a span modifies a noun, e.g., '*900mm wide*', then these would usually form two separate spans '*900mm*' and '*wide*'.

A second domain expert annotated 140 out of 200 sentences. The inter-annotator agreement was found to be Cohen k=0,79. We randomly divide the gold annotations into a 60/20/20 (%) split for train, development and test respectively – see Table 6 in Appendix A.

4 Training

4.1 Representing discontiguous spans

Regular tagging schemes, such as BIO and BI-OUL, are unable to represent discontiguous spans (Schneider et al., 2014). We adopt a tagging approach that can handle discontiguous pairs of spans, similar to (Muis and Lu, 2016). Specifically, we use BH to indicate the beginning (head) of a span and IH to indicate subsequent tokens that belong to this span. If a second span exists that is part of a discontiguous MWE, then the beginning (discontiguous) of this second span is tagged BD, and subsequent tokens of that span are tagged ID. Tags are provided with a type to distinguish between OBJECT and ACTION spans etc. Figure 2 provides example tags for an annotated sentence. Notably, we assume that between the head and discontiguous spans of a given type, there exist no head-spans of the same type.

4.2 Model

We adopt a sequence tagging approach that has been shown to work well (Huang et al., 2015), where an embedded text sequence is encoded by a bidirectional Long Short-Term Memory network
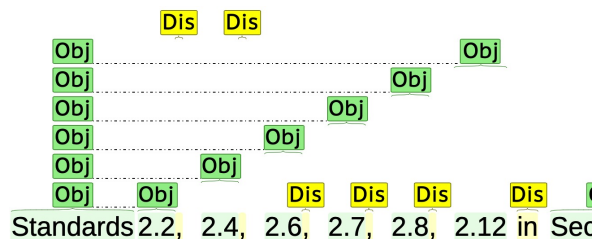

Figure 3: Example of an annotated coordination, where each of the items shares the word '*Standards*'. Note that only part of the sentence is shown here, the identifier for this sentence in the dataset is '*d_0.12.2_i3_#1_s_0*'.

(Hochreiter and Schmidhuber, 1997) and a Conditional Random Field (Lafferty et al., 1999) model is used to predict a tag for each token in the sequence. We modify the implementation found in AllenNLP[4] (Gardner et al., 2018). We embed text using pre-trained BERT embeddings (*bert-base-cased*) (Devlin et al., 2018) and rely on a single bi-LSTM layer (hidden dim. 384). We add a self-attention layer that captures relative positional information (Shaw et al., 2018), following (Huang et al., 2019). The encoded and attended representations are concatenated and projected through two linear layers (hidden dim. 60) before being passed to the CRF model. Table 2 provides an overview of results.

We compare the use of BERT (Devlin et al., 2018) and SPANBERT (*SpanBERT/spanbert-base-cased*) (Joshi et al., 2020) for tokenization and embedding. Our intuition is that the masked span Language Modelling task may help capture MWE properties used for MWE discovery, see 2.3. But in our experiments SPANBERT embeddings consistently perform worse, see Table 2. This may be due to a mismatch between the span types and sizes that SPANBERT was originally trained on – in SPANBERT all masked spans are contiguous – and the ones found in our training dataset. A more general issue is the imbalance of tag-types, see Table 3 for results on the test set per tag type.

5 Evaluation

Evaluating discovered lexical entries can be tricky (Constant et al., 2017). We limit our evaluation to the processing of OBJECT spans. First, we perform an extrinsic evaluation by comparing our model output against the defined terms in the SCOTREG

---

[4] https://github.com/allenai/allennlp-models

		P	R	F1
development	BERT	**80,18**	**81,76**	**80,96**
	SpanBERT	75,54	79,33	77,39
test	BERT	**79,93**	**80,89**	**79,93**
	SpanBERT	75,37	78,10	76,48

Table 2: Precision (P), recall (R), and F1-score (F1) on the development and test sets.

	P	R	F1	support
BH-obj	80,68	86,53	83,50	193
IH-obj	86,73	89,79	88,24	284
BD-obj	60,00	40,00	48,00	30
ID-obj	0,00	0,00	0,00	15
BH-act	76,15	83,84	79,81	99
IH-act	71,21	69,12	70,15	68
BD-act	50,00	37,50	42,86	8
ID-act	0,00	0,00	0,00	4
BH-dis	89,74	92,11	90,91	266
IH-dis	100,00	31,82	48,28	22
BH-func	65,62	63,64	64,62	66
IH-func	31,82	38,89	35,00	18
accuracy			80,89	1.073
macro avg.	59,33	52,77	54,28	1.073
weighted avg.	79,93	80,89	79,93	1.073

Table 3: Overview of results on the test set, broken down per tag type.

corpus. Second, post hoc human judgement provides insight in the number of correctly identified OBJECT spans given the sentence context.

5.1 Predicting defined terms

We use the model trained on SPAR.txt to predict tags for all sentences in the SCOTREG corpus. We then compare whether each of the 128 defined terms is identified at least once by our model. A total of 115 (89,84%) defined terms were found. Table 4 lists the 13 defined terms (10,16%) that were not found. Most of these defined terms exist verbatim in the SCOTREG corpus, but our model splits these spans into multiple parts. However:

- '*average flush*' only occurs as '*average flush volume*' and our model treats '*average*' as a separate FUNCTIONAL span.

- '*High-speed ready in-building physical infrastructure*' never occurs verbatim in the text.

5.2 Post hoc human judgement

We collect all contiguous and discontiguous OBJECT spans that our model predicted on the 13K sentences of the SCOTREG corpus. A total of 16,428 unique potential OBJECT spans are identified. We find that this number decreases slightly,

to 15,662, if we remove determiners and lower-case the text. We randomly select 165 out of the 16,5K OBJECT spans, and then one of the sentences in which this object occurs – this sample size provides a 99% confidence level with a 10% margin of error. We exclude objects that match any of the defined terms and exclude sentences that are part of the annotated dataset.

We use Doccano (Nakayama et al., 2018) for annotation. Each of these 165 samples is presented to the annotators as a combination of the OBJECT span and the corresponding sentence context. The task is to annotate whether the predicted OBJECT span is actually an object in the sentence, with a choice between the labels: (1) exact match, (2) partial match, or (3) not an object.

Two domain experts annotated the 165 OBJECT spans, see Table 5. The inter-annotator agreement was found to be Cohen k=0,79. If we take the average of their judgement, this comes down to 116 (70,3%) exact matches, 40 (23,6%) partial matches, and 9 (5,5%) non-objects. Examples of each labelled object and relevant parts of the sentence context include:

- **Exact match**: '*mechanical input air ventilation systems*' in "*Positive input systems - mechanical input air ventilation systems have been* [...]".

- **Exact match**: '*the warning*' in "[...] *the earliest possible warning* [...]" – where '*earliest*' and '*possible*' are modifiers that are not inherently part of the concept '*warning*'.

- **Exact match**: '*wall/roof junctions*' in "[...] *and at wall/roof junctions, wall/floor junctions and* [...]".

- **Partial match**: '*the control equipment*' in "[...] *the control and indicating equipment operates a fire alarm system* [...]" – '*the control and indicating equipment*' should be treated as a single OBJECT span.

- **Partial match**: '*Articles 15 & 16*' in "[...] *implements the terms of Articles 15 & 16 of Directive 2010/31/EU on* [...]" – should be split into '*Articles 15*' and '*Articles 16*'.

- **Partial match**: '*primary*' in "[...] *educations centres, schools (nursery, primary, secondary, special)* [...]" – should identify the discontiguous part '*school*'.

Undetected defined terms	What we extracted
Alternative exit	Probably split, we find '*alternative*' and '*exit*'
Average flush	Probably split, we find '*flush*' and several related terms, e.g., '*dual flush type*' and '*flush volume*'
Different occupation	Probably split, we find '*land*' and various types of '*occupation*', e.g., '*single occupation*', '*multiple occupation*', '*communal occupation*'
Land in different occupation	See above
Factory (class 1)	Probably split, we find '*factory*', '*factory building*' and their plurals
Factory (class 2)	See above, we do find '*'factory class 2'*'
High-speed ready in-building physical infrastructure	We find '*high-speed-ready in-building physical infrastructure*'
Major renovation works	Probably split, we find '*renovation*' and various types of '*works*', e.g., '*chemical works*', '*sewage work*' and '*protective works*'
Place of special fire risk	Probably split, we find the discontiguous span '*place*' '*fire risk*'
Public open space	Not found, we find various types of '*space*', e.g., '*clear space*', '*outdoor space*' and '*communal spaces*'
Reasonably practicable	Probably split, we find '*practicable*'
Storage building (class 1)	Probably split, we find '*storage building*' and '*storage buildings*'
Storage building (class 2)	See above

Table 4: Overview of the 13 defined terms that were not found in the text by our trained tagger.

| | Object | | |
	Exact match	Partial match	Not an object
Ann. 1	114 (69,1%)	41 (24,8%)	10 (6,1%)
Ann. 2	118 (71,5%)	39 (23,6%)	8 (4,8%)
Avg.	**116 (70,3%)**	**40 (24,2%)**	**9 (5,5%)**

Table 5: Overview of labels by two annotators on 165 OBJECT spans identified in the SCOTREG corpus.

- **Not an object**: '*sleeping*' in "*Rooms intended for sleeping should be* [...]" – should be an ACTION span here.

- **Not an object**: '*land subject*' in "[...] *development may be given approval on land subject to* [...]" – the OBJECT '*land*' is modified by '*subject to*'.

- **Not an object**: '*changes 1*' in "*Schedule 1 - changes to building types 1 and 20.*" – may be the result of overfitting on certain discontiguous patterns.

5.3 Discussion

Despite the small size of SPAR.txt, the trained model discovers a large number (16K) of OBJECT spans in 13K sentences. These spans cover most terms that are explicitly defined in the SCOTREG corpus (89,84%). The defined terms that the model did not identify are expressed in patterns that were never seen during training, although some do not occur verbatim in the texts. A significant proportion (70,3%) of identified OBJECT spans exactly match human judgement. Because annotation is cheap for our task, it is straightforward to create

additional gold training samples and improve performance. To this end, partial matches can help identify phenomena that were not seen during training, e.g., '*Articles 15 & 16*' and '*subject to*' as listed above. False positive OBJECT spans provide insight in phenomena that the model currently overfits on, and may potentially help balance future iterations of our dataset. Moreover, the predicted outputs for the SCOTREG corpus are valuable to the creation of a lexicon for ACC.

6 Conclusions

Regulatory documents are an important part of the legal framework, with research on ACC methods focusing on the grand goal of semantic parsing. This study introduces a much simpler parsing task that requires few training examples, with the additional benefit that the collection of a dataset is cheap. We presented the small SPAR.txt dataset and trained a sequence tagger that can process single-word and multi-word spans. We showed that the OBJECT spans identified in the SCOTREG corpus cover most of the existing, limited set of defined terms. Moreover, the model achieves reasonable accuracy when it comes to discovering OBJECT spans, regardless of whether these are discontiguous or not.

The annotation of gold training data for the presented approach is cheap, because the annotation task is simple. But the results can benefit the research on ACC, e.g., the output of our task may support more complicated semantic annotation tasks, IE and IR, as well as the development of a domain-specific lexicon. Future work will focus on clus-

tering the identified spans to develop a semantic lexicon, balancing and growing the dataset, as well as using predicted outputs for IR in support of manual ACC. Finally, we will explore how well the presented approach performs in other domains with similar text characteristics.

Acknowledgements

References

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly Supervised Learning of Semantic Parsers for Mapping Instructions to Actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, second edition, chapter 12, pages 267–292. Chapman and Hall.

Jakob Beetz, Jos Van Leeuwen, and Bauke De Vries. 2009. IfcOWL: A case of transforming EXPRESS schemas into ontologies. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, 23(1):89–101.

Phil Blunsom and Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *COLING/ACL 2006 - EMNLP 2006: 2006 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 164–171.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Nikolaos Aletras. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pages 2898–2904.

David L Chen. 2012. Fast online lexicon learning for grounded language acquisition. In *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, volume 1, pages 430–439.

Kenneth Church. 2013. How many multiword expressions do people know? *ACM Transactions on Speech and Language Processing*, 10(2):137–144.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword . . .*, June, pages 49–56.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, volume 1, pages 204–212. Association for Computational Linguistics.

Chris Cook. 2017. How legal drafting may be central to fire safety debate - BBC News.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arXiv preprint arXiv:1810.04805*.

Johannes Dimyadi and Robert Amor. 2013. Automated Building Code Compliance Checking. Where is it at? In *Proceedings of the CIB World Building Congress 2013 and Architectural Management & Integrated Design and Delivery Solutions (AMIDDS)*, 380, pages 172–185.

W. N. Francis and H. Kucera. 1964. A standard corpus of present-day edited American English, for use with digital computers. Technical report, Brown University, Providence.

G Frege. 1996. *Die Grundlagen der Arithmetik. Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Verlage Wilhelm Koebner, Breslau.

Stefan Fuchs. 2021. Natural Language Processing for Building Code Interpretation: Systematic Literature Review Report. Technical Report May, University of Auckland.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. *arXiv preprint arXiv:1803.07640*.

Spence Green, Marie Catherine De Marneffe, John Bauer, and Christopher D Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing TOUR DE FORCE with French. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 725–735.

Jonathan Herzig and Jonathan Berant. 2017. Neural semantic parsing over multiple knowledge-bases. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 2, pages 623–628. Association for Computational Linguistics (ACL).

Eilif Hjelseth and Nick Nisbet. 2011. Capturing Normative Constraints By Use of the Semantic Mark-Up Rase. In *Proceedings of CIB*, March, pages 26–28.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Nora Hollenstein, Nathan Schneider, and Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 3986–3990.

Cheng Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2019. Music transformer: Generating music with long-term structure. In *7th International Conference on Learning Representations, ICLR 2019*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991*.

Davies Hywel, Ant Burd, Rachel Smalley, Emma Clancy, Nigel Mason, Stephen Welch, Jeff House, Roger Holdsworth, Craig Renton, Rachel Cooper, Renford Gordon, Lorna Stimson, Graham Cleland, and Susan Johnson. 2020. Final Report of the Expert Group on Structure of Guidance to the Building Regulations. Technical report, Ministry of Housing, Communities & Local Government.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. In *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010*, pages 259–266.

John Lafferty, Andrew Mccallum, and Fernando Pereira. 1999. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data Abstract. *Departmental Papers (CIS)*, 2001(June):282–289.

Mirella Lapata and Alex Lascarides. 2003. Detecting Novel Compounds : The Role of Distributional Evidence. *Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics*, pages 235–242.

Sarah B Lawsky. 2017. Formalizing the Code. *SSRN Electronic Journal*, 377(17).

Kaijian Liu and Nora El-Gohary. 2017. Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Automation in Construction*, 81:313–327.

Christopher D. Manning, Hinrich Schütze, and Gerhard Weikurn. 2002. Foundations of Statistical Natural Language Processing. *SIGMOD Record*, 31(3):37–38.

Seonghyeon Moon, Gitaek Lee, Seokho Chi, and Hyunchul Oh. 2019. Automatic Review of Construction Specifications Using Natural Language Processing. In *Computing in Civil Engineering 2019*, pages 401–407, Reston, VA. American Society of Civil Engineers.

Seonghyeon Moon, Gitaek Lee, Seokho Chi, and Hyunchul Oh. 2021. Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing. *Journal of Construction Engineering and Management*, 147(1):04020147.

Raymond J Mooney. 2007. Learning for semantic parsing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4394 LNCS, pages 311–324. Springer.

Aldrian Obaja Muis and Wei Lu. 2016. Learning to recognize discontiguous entities. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 75–84.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text Annotation Tool for Human.

R. A. Niemeijer, B. De Vries, and J. Beetz. 2014. Freedom through constraints: User-oriented architectural design. *Advanced Engineering Informatics*, 28(1):28–36.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, pages 39–46.

Pieter Pauwels, Sijie Zhang, and Yong Cheol Lee. 2017. Semantic web technologies in AEC industry: A literature overview. *Automation in Construction*, 73:145–165.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *COLING/ACL 2006*, pages 651–658.

Marcos Pertierra, Sarah Lawsky, Erik Hemberg, and Una May O'Reilly. 2017. Towards formalizing statute law as default logic through automatic semantic parsing. In *CEUR Workshop Proceedings*, volume 2143.

138

Jim Plume and John Mitchell. 2007. Collaborative design using a shared IFC building model-Learning from experience. *Automation in Construction*, 16(1):28–36.

Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and Newton Howard. 2014. Dependency-Based Semantic Parsing for Concept-Level Text Analysis. In *CICLing*, pages 113–127.

Cornelius Preidel and André Borrmann. 2018. BIM-based code compliance checking. In *Building Information Modeling: Technology Foundations and Industry Practice*, pages 367–381. Springer International Publishing.

Dheeraj Rajagopal, Erik Cambria, Daniel Olsher, and Kenneth Kwok. 2013. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, pages 565–570.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaite, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang Qasem-iZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the Parseme shared task on automatic identification of verbal multiword expressions. In *LAW-MWE-CxG 2018 - Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions, Proceedings of the Workshop*, pages 222–240.

Martin Riedl and Chris Biemann. 2015. A single word is not enough: Ranking multiword expressions using distributional semantics. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 2430–2440.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 2276, pages 1–15.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A Smith. 2014. Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Scottish Government. 2020. Building Standards Technical handbook 2020: Domestic.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 2, pages 464–468.

Jeffrey Mark Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2 SPEC. ISS.):39–91.

Wawan Solihin, Johannes Dimyadi, and Yong-Cheol Lee. 2019. In Search of Open and Practical Language-Driven BIM-Based Automated Rule Checking Systems. In *Advances in Informatics and Computing in Civil and Construction Engineering*, pages 577–584. Springer International Publishing.

Robert Speer, Catherine Havasi, R Speer, and C Havasi. 2013. ConceptNet 5: A Large Semantic Network for Relational Knowledge. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, chapter 6, pages 161–176. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *13th Conference of the European Chapter of the ACL*, Figure 1, pages 102–107.

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions.

Adam Wyner, Johan Bos, Valerio Basile, and Paulo Quaresma. 2012. An empirical approach to the semantic representation of laws. In *Frontiers in Artificial Intelligence and Applications*, volume 250, pages 177–180.

Xin Xu and Hubo Cai. 2019. Semantic Frame-Based Information Extraction from Utility Regulatory Documents to Support Compliance Checking. In *Advances in Informatics and Computing in Civil and Construction Engineering*, pages 223–230. Springer International Publishing, Cham.

Luke S Zettlemoyer and Collins Michael. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*, pages 658–666.

Jiansong Zhang and Nora M. El-Gohary. 2016a. Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking. *Automation in Construction*, 73:45–57.

Jiansong Zhang and Nora M. El-Gohary. 2016b. Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *Journal of Computing in Civil Engineering*, 30(2):04015014.

Ruichuan Zhang and N. El-Gohary. 2019. A machine learning-based method for building code requirement hierarchy extraction. *Proceedings, Annual Conference - Canadian Society for Civil Engineering*, 2019-June:1–10.

Ruichuan Zhang and El-Gohary Nora. 2020. A Machine-Learning Approach for Semantically-Enriched Building-Code Sentence Generation for Automatic Semantic Analysis. In *Construction Research Congress*, pages 1261–1270.

Botao Zhong, Xuejiao Xing, Hanbin Luo, Qirui Zhou, Heng Li, Timothy Rose, and Weili Fang. 2020. Deep learning-based extraction of construction procedural constraints from construction regulations. *Advanced Engineering Informatics*, 43(December 2019):101003.

Peng Zhou and Nora El-Gohary. 2019. Text and Information Analytics for Fully Automated Energy Code Checking. In *Sustainable Civil Infrastructures*, pages 196–208. Springer Science and Business Media B.V.

A Overview of spans and tags for SPAR.txt and SCOTREG predictions

	All	**Train**	*%**	**Development**	*%**	**Test**	*%**	**Predicted**
Domestic	98	62	63,27	19	19,39	17	17,35	6.313
Non-domestic	102	58	56,86	21	20,59	23	22,55	7.293
								13.606
avg. sent. (token†) length	27,96	28,31		28,05		26,83		26,72
Std. dev	15,31	15,50		14,03		15,92		15,53
Shortest	2	2		4		2		1
Longest	85	85		60		63		199
Span types	3.253	1.985	61,02	633	19,46	635	19,52	216.949
Discourse	1.312	788	60,06	258	19,66	266	20,27	87.957
Object	1.122	695	61,94	224	19,96	203	18,09	74.193
Action	476	294	61,76	82	17,23	100	21,01	31.835
Functional	343	206	60,64	69	20,12	66	19,24	22.964
Nr. discontiguous	193	122	63,21	33	17,10	30	15,54	6,450 Objects 2,007 Actions
avg. span (char) length	7,51	7,49		7,49		7,58		7,14
Discourse	2,56	2,50		2,53		2,77		2,44
Object	12,54	12,41		12,40		13,11		12,40
Action	9,51	9,54		9,87		9,15		8,31
Functional	7,22	7,08		7,29		7,58		6,53
Tag types								
IH-obj	1.551	931	60,03	336	21,66	284	18,31	108.507
BH-dis	1.312	788	60,06	258	19,66	266	20,27	87.957
BH-obj	1.071	659	61,53	219	20,45	193	18,02	74.311
BH-act	472	292	61,86	81	17,16	99	20,97	31.843
BH-func	343	208	60,64	69	20,12	66	19,24	22.964
IH-act	331	205	61,93	58	17,52	68	20,54	19.013
BD-obj	149	94	63,09	25	16,78	30	20,13	6.450
IH-func	136	88	64,71	30	22,06	18	13,24	6.784
ID-obj	134	88	65,67	31	23,13	15	11,19	3.114
IH-dis	45	16	35,56	7	15,56	22	48,89	482
BD-act	34	20	58,82	6	17,65	8	23,53	2.007
ID-act	14	8	57,14	2	14,29	4	28,57	74

Table 6: Overview of sentence and span statistics for SPAR.txt (Gold/Train/Dev/Test), as well as for the predictions (Predictions) over the entire SCOTREG corpus. %* indicates the percentage of all gold data. † Note that BERT tokenization is used here.

B Overview of examples

Sect.	Full sentence	Source	
3.2	**A roof covering or roof light which forms part of an internal ceiling lining should** also follow the guidance to Standard 2.5 Internal linings.	2.8.0 Introduction	Domestic
3.3	**The Building (Scotland) Act 2003 gives Scottish Ministers** the power to make building regulations to:	0.1.1 Introduction	Domestic
	In order to allow unobstructed access to a domestic building for fire and rescue service personnel, **a paved (or equivalent) footpath at least 900mm wide** (see also Section 4 Safety) should be provided to the normal entrances, of a building.	2.12.4 Access for fire and rescue service personnel	Domestic
3.4	Section 6 Energy, indicates that less demanding U-values can be adopted for the **insulation envelope** of certain types of limited life buildings, other than dwellings and residential buildings.	0.6.1 Explanation	Domestic
	In conversions for example, it may be easier to install **self-contained emergency luminaries** than to install a protected circuit to the existing lighting system	2.10.2 Protected circuits	Domestic
3.5	To enable the continued use of existing stocks of building modules and sub-assemblies, subject to fabric insulation meeting the U-values noted in clause 6.C.3, a modifying factor can be applied to increase **the Target Emissions Rating** (TER) for the building.	n_6.C257_i0_s_0	Non-domestic
	This is termed '**offensive fire-fighting**' and is normal practice regardless of whether people are in the building or not.	d_2.14.0_i4_s_1	Domestic
	BS 8000-15: 1990 - Workmanship on building sites	d_0.8.8_i3_#1_s_38	Domestic
	The minimum 3m separation in the diagram below **should be maintained** between each $5m^2$ panel.	d_2.5.7_i5_#1_s_0	Domestic
	The guidance in this clause **takes account of** the audibility levels in adjoining rooms and **the effect of smoke travelling along a ceiling.**	d_2.11.7_i0_s_0	Domestic
	The guidance in this clause **takes account of** the audibility levels in adjoining rooms and the effect of smoke travelling along a ceiling.	d_2.11.7_i0_s_0	Domestic
	CO2 monitoring equipment should be provided in the apartment expected to be **the main or principal bedroom** in a dwelling where infiltrating air rates are less than $15m^3$/hr/m^2 @ 50 Pa.	d_3.14.2_i1_s_0	Domestic
	Non-domestic use within dwellings - accommodation **up to** $50m^2$ used by an occupant of a dwelling in their professional or business capacity should be considered as a part of the dwelling.	d_6.9.1_i3_s_0	Domestic
	In the measurement of height or depth from ground which **is not level** the height or depth shall be taken to be the mean height or depth, except that: for the purpose of types 1, 2, 3, 4, 5, 18 or 19 of schedule 3, and for any other purpose where the difference in level is more than 2.5m the height or depth shall be taken to be the greatest height or depth.	d_0.7.2_i1_#4_s_0	Domestic
	Other pipes should be capped at both ends and at any point of connection, to ensure rats **cannot gain** entry.	n_3.5.5_i1_s_1	Non-domestic
	In order to allow unobstructed access to a domestic building for **fire and rescue service personnel**, a paved (or equivalent) footpath at least **900mm wide** (see also Section 4 Safety) should be provided to the normal entrances, of a building.	d_2.12.4_i1_s_0	Domestic
	The layout of a dwelling, **the size and orientation of the windows**, the thermal mass, level of insulation, airtightness, and ventilation can have a significant affect on the demand for heat.	d_3.13.1_i0_s_0	Domestic
	Standard 7.1 - amendments have been made to guidance with regard to the carbon dioxide (CO2) emissions target within **the Silver and Gold level** of Sustainability labelling in relation to the CO2 emissions target introduced by the 2015 energy standards.	d_7.0.5_i1_#0_s_0	Domestic

Sect.	Full sentence	Source	
	Materials that are susceptible to changes in their properties may be used in building work and will meet the requirements of the regulations if the residual properties, including **the structural properties**:	d_0.8.7_i1_s_0	Domestic
	The collation and dissemination of information relating to **matters of structural concern** is a vital element of achieving safe structures.	d_1.0.1_i2_s_2	Domestic
5.2	Positive input systems - **mechanical input air ventilation systems** have been successfully installed in existing dwellingswith the objective of overcoming problems of surface condensation and mould growth.	3.14.11 Mechanical ventilation and systems	Domestic
	This is to give occupants and staff **the earliest possible warning** of an outbreak of fire and allow time for assisting occupants in an emergency to evacuate the building or for horizontal progressive evacuation initially to an adjacent sub-compartment which leads to a compartment exit.	2.11.5 Hospitals	Non-domestic
	These 'bridges' commonly occur around openings such as lintels, jambs and sills and at **wall/roof junctions**, wall/floor junctions and where internal walls penetrate the outer fabric.	3.15.4 Surface condensation - thermal bridging	Non-domestic
	Normally **the control and indicating equipment** operates a fire alarm system and it may perform other signalling or control functions as well.	2.11.3 Categories of fire detection and fire alarm system	Non-domestic
	This regulation implements the terms of **Articles 15 & 16** of Directive 2010/31/EU on the Energy Performance of Buildings (EPBD).	0.17.1 Explanation	Domestic
	• education centres, **schools (nursery, primary, secondary, special)**	6.9.3 Location of an energy performance certificate	Non-domestic
	Rooms intended for **sleeping** should be separated by a door that will act as a sound barrier and reduce noise transference.	5.2.5 Doors in internal walls	Non-domestic
	Pressure for land development may mean that development may be given planning approval on **land subject** to some risk of flooding.	3.3.0 Introduction	Domestic
	• Schedule 1 - **changes to building types 1** and 20.	0.2.1 Explanation of Regulation 1	Non-domestic

Table 7: Complete overview of the examples used in the paper and their location in SCOTREG. Examples that occur in SPAR.txt have an identifier, such as n_6.C257_i0_s_0. These identifiers are created as follows: domestic (d) or non-domestic (n), the section (6.C2) – in this case concatenated with an integer (57) that is incremented each time a section name is reused, the item index (0 for first item) that indicates which item in the list of this section, and the sentence number (s_0 for first sentence). '#' in the identifier means that the sentence occurs in a list or table, with lists and tables 0-indexed for each section.