Challenges in Technical Regulatory Text Variation Detection

Shriya Vaagdevi Chikati [a,c], **Samuel Larkin** [a], **David Minicola**[b], **Chi-kiu Lo 羅致翹**[a*]
[a]Digital Technologies Research Centre [b]Construction Research Centre
National Research Council Canada
[c]University of Waterloo
chikatishriya@gmail.com, {Samuel.Larkin, David.Minicola, Chikiu.Lo}@nrc-cnrc.gc.ca
[*]corresponding author

Abstract

We present a preliminary study on the feasibility of using current natural language processing techniques to detect variations between the construction codes of different jurisdictions. We formulate the task as a sentence alignment problem and evaluate various sentence representation models for their performance in this task. Our results show that task-specific trained embeddings perform marginally better than other models, but the overall accuracy remains a challenge. We also show that domain-specific fine-tuning hurts the task performance. The results highlight the challenges of developing NLP applications for technical regulatory texts.

1 Introduction

In Canada, the regulation of building construction is the jurisdiction of the provinces and territories (P/T). National Research Council of Canada is responsible for publishing the National Model Construction Codes[1] (NMCC), a collection of model construction codes, to help promote consistency among the P/T construction codes. The model construction codes set out the technical requirements for the design and construction of new buildings, as well as the change of use and demolition of existing buildings. The NMCC are adopted individually by each P/T legislatures while modifications are made to suit local needs. As a result, there are various kinds of variations between the NMCC and each of the P/T construction codes. Variations in construction codes create barriers to the free movement of construction goods, services, and investments within Canada. To support inter-P/T trades and the mobility of talents, harmonizing the construction codes across Canada is a key priority in the development and maintenance of the codes and tracking code variations is the first step towards code harmonization. Due to the high volume and high technicality of the construction codes, tracking variations in them is a difficult and expensive task that requires labor-intensive studies done by codes advisors. Some of the challenges in technical domain knowledge include identifying similarity/dissimilarity in definitions of technical terms or material specifications when they are applied in context, e.g. “sawdust” could be equivalent to “combustible pallets” in the fire codes. This study explores the potential of leveraging recent development of natural language processing (NLP) techniques to detect variations between the NMCC and the adopted codes in each P/T.

The NMCC consists of the National Building Code (NBC), the National Fire Code (NFC), the National Plumbing Code (NPC), and the National Energy Code for Buildings(NECB). Each P/T has the autonomy in deciding how the NMCC are adapted, or referenced. For example, for the 2015 edition of NBC, NFC, and NPC, New Brunswick, Nova Scotia, Manitoba and Saskatchewan adopted them with some modifications and additions; while Ontario developed their codes based on the national models, but with significant variations in content and scope and made major reorganization of the order of the provisions. The technical provisions of the construction codes (both the NMCC and the P/T construction codes) are structured, segmented and numbered and the detailed provisions are found at the sentence level. Hence, the majority of variations tracked at the sentence level. The variations in the construction codes are classified by National and P/T codes advisors into the following categories:

1. **P/T Only (Addition)** - The P/T has decided to add a new technical provision (i.e. a sentence) to their local construction codes.

2. **National Only (Deletion)** - The P/T has decided not to include a technical provision of

[1] https://nrc.canada.ca/en/certifications-evaluations-standards/codes-canada/codes-canada-publications

Proceedings of the 31st International Conference on Computational Linguistics, pages 5–9
January 19–24, 2025. ©COLING 2025

the NMCC in their local construction codes.

3. **Common Sentence** - The technical provision appear in both the the NMCC and the corresponding local construction codes. This category is further divided into two subcategories: a) adopted without changes or with editorial changes and b) modified with technical variations. Table 2 shows some examples of sentences with editorial changes vs. technical variations.

The sentences in the second example of editorial changes in table 2 of Appendix A are rewritten with a different focus, although remain technically the same. On the other hand, for the first example of technical variations, the sentences are nearly identical, except for the specified distance. This presents a huge challenge in classifying the sentences into the two subcategories because comprehensive technical knowledge is required to identify the technical variations. Therefore, in this preliminary study, we focus on the three main categories (i.e. P/T Only, National Only and Common Sentence) and we then formulate the task of variations classification as a sentence alignment task. “Common Sentence” in nation codes and P/T codes should be aligned while “P/T Only” and “National Only” should remain unaligned. The challenge here, though, is that the sentences in the two documents are not in the same order and we show that off-the-shelf sentence alignment tools perform poorly in this task.

In this paper, we study the potential and challenges of detecting variations between the NMCC and the construction codes in each P/T. We evaluate the performance of different sentence representation models on the sentence alignment task. Our results show that developing NLP applications for technical regulatory texts requires more investments and efforts.

2 Sentence Alignment Methods

Sentence alignment methods have two major components: 1) the sentence representation model for reflecting how similar sentences are to each other; and 2) alignment extraction approach for deciding the sentence alignment given the sentence similarity scores.

2.1 Sentence Representation

Representing sentences in a vector space enables numeric computation of their relations, such as similarity, which then enables practical tasks, like text classification and information retrieval. Similarity of sentences in vector space is usually computed using cosine similarity that measures the angular distance of their vectors. There are two generations of sentence representation models, distributional and neural.

2.1.1 Sparse distributional Vector Model

The most commonly used distributional text representation is the bag-of-words model (BoW). The BoW model considers a text as an unordered collection of words and the frequency of occurrence of each word in the text is the value of the dimension representing that word. Variants of the BoW model include 1-hot encoding that only considers the presence of each word rather than their frequency; and term frequency-inverse document frequency (tf-idf) weighted BoW that weights the frequency of each word by their tf-idf. Both variants intend to down-weight the influence of function words (more frequently occurring) and up-weight the influence of content words (less frequently occurring) in representing the text. Sparse distributional vector models suffer from the curse of dimensionality as the dimensions of the vector space grows with the size of the vocabulary of the sentences it represents.

2.1.2 Neural Sentence Embeddings

In contrast with distributional vector model, neural sentence embeddings are learned representation models. LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2022) are the most commonly used massively multilingual sentence embeddings models. While LASER and LaBSE are pretrained for general purpose sentence representation, some of the sentence embedding models are further trained for specific tasks to obtain better performance. We choose to experiment with a few sentence embedding models that perform well on the English semantic textual similarity task because our sentence alignment task relies on the sentence similarity scores. The models are bilingual-embedding-large (Javaness, 2024), multilingual-e5-large-instruct (Wang et al., 2024), mxbai-embed-large-v1 (Shakir et al., 2024) and GIST-embedding-v0 (Solatorio, 2024).

2.2 Alignment Extraction

State-of-the-art parallel text sentence alignment algorithms, such as Vecalign (Thompson and Koehn, 2019) and SentAlign (Steingrimsson et al., 2023), take advantage of parallel text where sentences in the two documents are roughly in monotonic order with local sentence reordering and use approximation or divide and conquer approaches to reduce the time and space complexity of the alignment extraction component in the algorithm. However, as we mentioned before, the P/T construction codes have major reorganization of the order of the provisions when comparing to the NMCC such that the assumption of monotonic sentence order does not hold for the task at hand.

The alignment extraction component is naturally a maximum weight bipartite matching problem. We consider all the sentences in one document as one partition of vertices in a graph, all the sentences in the other document as the other partition of vertices of the graph and the sentence similarity scores between all the sentence pairs as the weights of the edges. The matching produced by the maximum weight matching algorithm, that maximizes the sum of the weights of the edges included in the matching, is then the optimal sentence alignment. Thus, we propose to use the Scipy library[2] implementation of the LAPJVsp algorithm (Jonker and Volgenant, 1987) for the maximum weight matching problem in our experiment.

3 Experiments

The NMCC are updated and published once every five years and each of the P/T adopts the updated version in subsequent years. National and P/T codes advisors then work on analyzing the variations. The latest edition of NMCC with all variations tracked and classified was published in 2015 (Canadian Commission On Building And Fire Codes, 2015a,b,c; Canadian Commission On Building And Fire Codes and Natural Resources Canada, 2015). In our experiments, we use the English version of the 2015 edition of the NMCC and the corresponding P/T construction codes. We create the training and testing sets for our experiments by doing a random 80/20 split.

3.1 Evaluation Metric

To evaluate the performance of the construction code variation detection, we use the alignment error rate (AER) of the predicted alignment (Och and Ney, 2000). In our task, we only have the set of "sure" gold standard alignment (i.e. the Common Sentence labeled by domain experts) but not the set of possible alignment. The AER is, therefore, simplified as $AER = 1 - \left(\frac{2 \times |P \cap G|}{|P| + |G|}\right)$ where P is the set of predicted alignment and G is the set of gold standard alignment. The lower AER means that the alignment model is better.

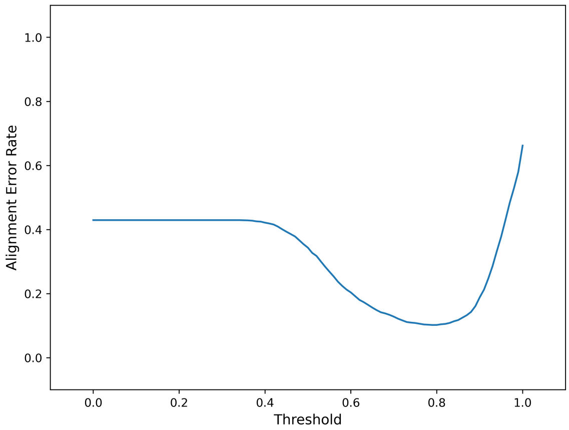

Figure 1: Training AER of the step-wise search for the optimal similarity score threshold for determining aligned sentence pairs.

3.2 Solution to Over-Alignment

The problem of using maximum weight matching to extract aligned sentence pairs is that it always returns a full matching such that every sentence in the smaller set is matched with a sentence in the other set. That means the extracted alignment would be over-aligned, in the sense that a sentence pair would be returned as aligned even with low similarity score. To reclassify the over-aligned sentence pairs into "P/T Only" and "National Only", we do a step-wise search on the training data to determine a threshold of similarity score for sentence pairs to remain aligned. As we increase the similarity score threshold for sentence pairs to remain aligned by 0.01 at each step, the AER drops until the threshold is too high and starts separating sentence pairs that ought to stay aligned to each other. This is being done for each individual model, including the domain fine-tuned model. Figure 1 shows a typical plot of the training AER against the similarity score threshold in the step-wise search. We then apply the optimal threshold that results in the lowest training AER to the test set.

3.3 Domain Specific Fine-tuning

Since the technical provisions in the construction codes are highly domain specific, we also experiment with fine-tuning the sentence representation model using contrastive training with the in-domain

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csgraph.min_weight_full_bipartite_matching.html

training data. We choose to fine-tune the GIST-Embedding-v0 model because it achieves the lowest training AER. Contrastive training leverages positive and negative sentence pairs to teach the model how to differentiate between similar and dissimilar sentences. This process helps the model to push dissimilar sentences further apart and pull similar sentences closer together in the embedding space. In our experiment, the positive pairs are the “Common Sentence” and we create the negative pairs by pairing the “National Only” sentences with the “P/T Only” sentences. Thus, the size of the fine-tuning data set is the same as the training data set, i.e. 80% of the complete NMCC, around four thousand sentence pairs. We used the default learning rate in fine-tuning.

4 Results

Table 1 shows the training and testing AER of the experiments on using different sentence representation models for detecting construction code variations. We see that the parallel text alignment baseline, Vecalign, is clearly not suitable for our task. Vecalign uses LASER as the underlying sentence representation model and assumes the sentences in the two input documents are in monotonic order. When we compare the performance of Vecalign with that of our experiment on the LASER model, we conclude that our proposal of using maximum weight matching algorithm for alignment extraction is more suitable for the task.

Model	Train	Test
parallel text alignment baseline		
Vecalign	0.4564	0.4568
distributional vector-based		
Bag-of-Words (BoW)	0.1402	0.1554
1-hot	0.1296	0.1426
tf-idf weighted BoW	0.1233	0.1372
pretrained sentence embeddings		
LASER	0.1635	0.1783
LaBSE	0.1352	0.1471
task specific trained embeddings		
bilingual-embedding-large	0.1183	**0.1306**
multilingual-e5-large-instruct	0.1210	0.1403
mxbai-embed-large-v1	0.1194	0.1366
GIST-embedding-v0	**0.1165**	0.1339
construction codes fine-tuned		
GIST-embedding-v0	0.1370	0.1522

Table 1: Training and testing AER of the experiments on using different sentence representation models for detecting construction code variations.

The tf-idf weighted bag-of-words model performs better than both of the pretrained sentence embedding models (LASER and LaBSE) and only marginally worse than the task specific trained embedding models.

The domain specific fine-tuning model performs significantly worse than the base model before fine-tuning. This is perhaps due to limited amount of in-domain data used in fine-tuning the model.

Overall, the AER for all our experiments are high. With an AER higher than 13% by all models before attempting to classify editorial and technical variation, we demonstrated that NLP research and development on technical regulatory texts remains an open question and great challenge.

5 Conclusion

In this study, we explored the potential of using current natural language processing techniques to detect variations between the National Model Construction Codes and the construction codes of different Canadian P/T. We evaluated various sentence representation models on this task. The overall bad performance across all models suggests that current NLP technologies are not yet fully equipped to handle the complexity of technical regulatory text. This highlights the need for further research in developing NLP models that could acquire the necessary technical knowledge from technical regulatory text and improving the accuracy and reliability of NLP tools in technical regulatory application.

Limitations and Ethical Considerations

Our work on automatic construction codes variation detection is intended to assist National and P/T codes advisors in tracking and analyzing variations, with the goal of harmonizing codes across Canada. The output of automatic construction codes variation detection *will NOT and should NOT* be used directly by any code users or code enforcement bodies before verification by technical experts in Canadian construction codes. As the accuracy of the current experiments are low, the risk of resulting in misleading information is high if the model output is directly used by code users or enforcement bodies. Misusing the model output could lead to financial loss, noncompliance or wrongful enforcement of construction codes. We think that it is of utmost importance to restrict the use of the model output to technical experts specialized in construction code variation identification.

Code	Sentence
Common Sentence (editorial)	
NBC	Where a fire safety plan is required, it shall conform to Section 2.8. of Division B of the NFC.
British Columbia BC	Fire safety plans shall conform to the British Columbia Fire Code.
NBC	Visual signal devices required by Sentence 3.2.4.19.(1) shall continue to emit a visible signal while voice instructions are being transmitted.
Ontario BC	The voice communication system referred to in Sentence (1) shall be designed so that visual signal devices are not interrupted while voice instructions are being transmitted.
Common Sentence (technical)	
NBC	The developed length of a building sewer between the building and the first manhole to which the building sewer connects shall not exceed 75 m.
Ontario BC	The developed length of a building sewer between the building and the first manhole to which the building sewer connects shall not exceed 30 m.
NFC	The removal, abandonment in place, disposal or temporary taking out of service of an underground piping system shall be in conformance with good engineering practice. (See Note A-4.3.16.1.(1).)
Alberta FC	Corrosion protection systems shall be maintained in operating condition when a storage tank is temporarily out of service and during seasonal shutdowns.

Table 2: Examples of sentences in the NMCC modified with editorial changes vs. technical variations in the P/T construction codes.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Canadian Commission On Building And Fire Codes. 2015a. National building code of canada: 2015.

Canadian Commission On Building And Fire Codes. 2015b. National fire code of canada: 2015.

Canadian Commission On Building And Fire Codes. 2015c. National plumbing code of canada: 2015.

Canadian Commission On Building And Fire Codes and Natural Resources Canada. 2015. National energy code of canada for buildings: 2015.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

La Javaness. 2024. Bilingual-embedding-large.

R. Jonker and A. Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Aamir Shakir, Darius Koenig, Julius Lipp, and Sean Lee. 2024. Boost your search with the crispy mixedbread rerank models.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *Preprint*, arXiv:2402.16829.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

A Examples of Common Sentence

Table 2 shows the example of aligned sentences in NMCC and P/T construction codes with different variation types (editorial vs. technical).