

# Bootstrapping Supervised Product Taxonomy Mapping with Hierarchical Path Translations for the Regulatory Intelligence Domain

Alfredo Maldonado<sup>1</sup>, Spencer Sharpe<sup>2</sup>, and Paul ter Horst<sup>3</sup>

<sup>1,3</sup> UL, Dublin, Ireland

<sup>2</sup> UL, Laramie, WY, United States

{Alfredo.Maldonado, Spencer.Sharpe, Paul.TerHorst}@ul.com

## 1 Introduction

Regulatory Intelligence (RI) help manufacturers and retailers understand their compliance requirements in the markets they intend to serve. Automatic RI relies on matching product taxonomies with regulations. However, retailers use an abundance of different product taxonomies. This work addresses this problem by investigating automatic entity alignment between arbitrary vendor-specific taxonomies to the GS1's Global Product Classification (GPC)<sup>1</sup>.

Taxonomy mapping usually takes place at the beginning of the onboarding process. This means that no historic alignments are available, limiting the applicability of supervised mapping methods, unless we have a reliable seeding method to use in lieu of historic alignments. We describe such a seeding approach and use it to train a supervised neural mapping system. The seeding approach is inspired in neural machine translation. Crucial in the approach is the hierarchical classification [1–3], where a hierarchical class is represented as a sequence of node IDs. In the GPC product taxonomy, for example, a text label “powered stationary exercise bicycle” has a node ID of 10005815 – Cycles (Powered), and a full taxonomic classification in the branch [71000000, 71010000, 71010800, 10005815]. We train a sequence-to-sequence architecture with attention [4] on examples of this sort, mapping the text label to the sequence of node IDs. Training data was acquired from GPC and a small set of product names (~1,200) manually labelled with brick codes. The resulting model allows us to predict GPC mappings for a target taxonomy. The only requirement is that both taxonomies are expressed in the same human language. However, equivalent labels in the two taxonomies need not be textually identical due to word embeddings.

In the second stage, the seeding from the seq2seq model is used as training data for Deep Graph Matching Consensus (DGMC) [5], which learns mappings in two steps: first, it learns links between two graphs via localised node embeddings. Then, it refines these initial correspondences via neighbourhood consensus.

---

<sup>1</sup> <https://www.gs1.org/standards/gpc>

## 2 Evaluation

Our experiments were run on the WDC Product Categorisation Gold Standard<sup>2</sup>, which links the Google Product Taxonomy to GPC. Our system (Seq2Seq-DGMC in Table 1) takes the Google Taxonomy and the GPC Taxonomy as input, along with a set of seed mappings (predicted by Seq2Seq) in order to train the DGMC model. The trained model outputs suggested mappings between the two taxonomies. Table 1 compares our system with the supervised alternatives:

- Supervised-DGMC50: DGMC system trained on a 50% sample of the Google-GPC mappings and evaluated on the other 50%.
- Supervised-DGMC10: DGMC system trained on a 10% of the Google-GPC mappings and evaluated on the other 90%.
- Seq2Seq: The raw output of our seeding method.

**Table 1.** Experiment Results on different system configurations using the “Hits at 1” (H@1) and “Hits at 10” (H@10) metrics commonly used for ranking systems.

System	H@1	H@10	System	H@1	H@10
<i>Supervised-DGMC50</i>	.47	.78	Seq2Seq	.26	N/A
<i>Supervised-DGMC10</i>	.31	.68	Seq2Seq-DGMC	.31	.63

The performance of our automatically seeded Seq2Seq-DGMC taxonomy mapping is comparable to that of the manually seeded Supervised-DGMC10. Depending on the sizes of the taxonomies to map, this result translates to potentially significant time savings. As further work, we will evaluate this method with more vendor-specific product taxonomies. We will also directly measure time saved in manually correcting automatic mappings. We also wish to explore additional taxonomy matching models.

## References

1. Yang, Z., Liu, G.: Hierarchical Sequence-to-Sequence Model for Multi-Label Text Classification. *IEEE Access*. 7, 153012–153020 (2019).
2. Umaashankar, V., Shanmugam S., G.: Multi-Label Multi-Class Hierarchical Classification using Convolutional Seq2Seq. In: KONVENS (2019).
3. Hasson, I., Novgorodov, S., Fuchs, G., Acriche, Y.: Category Recognition in E-Commerce using Sequence-to-Sequence Hierarchical Classification. In: ACM International Conference on Web Search and Data Mining (2021).
4. Dzmitry Bahdana, Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation By Jointly Learning To Align and Translate. In: ICLR (2015).
5. Fey, M., Lenssen, J.E., Morris, C., Masci, J., Kriege, N.M.: Deep Graph Matching Consensus. In: ICLR (2020).

<sup>2</sup> <http://webdatacommons.org/structureddata/2014-12/products/gs.html#toc4>