Retrieving Music Semantics from Optical Music Recognition by Machine Translation

Martha E. Thomae[1], Antonio Ríos-Vila[2], Jorge Calvo-Zaragoza[2], David Rizo[2], and José M. Iñesta[2]

[1] *Schulich School of Music, McGill University*
[2] *Department of Software and Computing Systems, University of Alicante*

Introduction

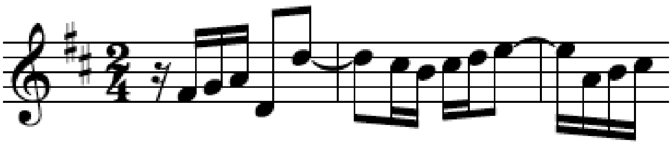
We present a machine learning-based approach to retrieve the semantics of a sequence of (graphic) music symbols, which constitutes a central problem in the field of Optical Music Recognition (OMR). OMR is the process of converting the digital image of a score into a symbolic file encoding the music content of that score. The traditional OMR workflow consists of four stages: preprocessing, symbol recognition, music reconstruction, and music encoding. The third stage, music reconstruction, must retrieve the actual musical meaning of the graphical symbols recognized in the previous stage. So far, the models proposed to solve this problem are based on rules (Rossant and Bloch, 2006) or grammars (Couasnon, 2001; Szwoch, 2007), which prevents their use in other contexts (e.g., in notation systems other than the one for which they were implemented). For high scalability, we propose a machine learning-based approach which learns the semantics of a particular notation system by providing the model with enough training examples.

We use an encoding introduced by Calvo-Zaragoza and Rizo (2018b) to represent the graphical and semantic information obtained by the second and third stages of the OMR workflow, respectively. This encoding provides an intermediate representation that, during the music encoding stage of the OMR process, becomes a well-established music format, such as MusicXML, MEI, or **kern.

Agnostic and Semantic Encodings of Music Sequences

In MEC'2017, Rizo et al. (2017) presented the concept of agnostic and semantic sequential representations of a music score. The agnostic encoding represents the output of the music symbol recognition stage of the OMR, where we only have the graphical information about the symbols (their shapes and positions) and no musical meaning. The agnostic representation is a sequential encoding of the graphical symbols in the score (Figure 1b). Each token in the sequence encodes two types of information: the label of the symbol (e.g., C clef, quarter note, half note, sharp) and its vertical position within the staff (e.g., third line, fourth space). On the other hand, the semantic representation is a sequential encoding of symbols in a score, which includes their musical meaning (Figure 1c). Translating an agnostic sequence into a semantic one involves several tasks, including re-interpreting a series of accidentals into a key signature and parsing the position of the notes in the staff into pitch values.

Calvo-Zaragoza and Rizo (2018b) showed that sequential encodings are suitable for converting a digital image into either an agnostic or a semantic representation without human-encoded rules, with more robust results in the agnostic case (Calvo-Zaragoza and Rizo, 2018a). In this paper, we implement a machine translator that takes the agnostic representation of a sequence of notes in the staff and generates its corresponding semantic representation, in order to take advantage of the performance of the agnostic case for OMR.



(a) Music excerpt.

```
clef.G-L2, accidental.sharp-L5, accidental.sharp-S3, digit.2-L4, digit.4-L2, rest.sixteenth-L3,
note.beamedRight2-S1, note.beamedBoth2-L2, note.beamedLeft2-S2, note.beamedRight1-S0,
note.beamedLeft1-L4, slur.start-L4, barline-L1, slur.end-L4, note.beamedRight1-L4,
note.beamedBoth2-S3, note.beamedLeft2-L3, note.beamedRight2-S3, note.beamedBoth2-L4,
note.beamedLeft1-S4, slur.start-S4, barline-L1, slur.end-S4, note.beamedRight2-S4,
note.beamedBoth2-S2, note.beamedBoth2-L3, note.beamedLeft2-S3
```

(b) Agnostic encoding of the music excerpt.

```
clef-G2, keySignature-DM, timeSignature-2/4, rest-sixteenth, note-F#4_sixteenth,
note-G4_sixteenth, note-A4_sixteenth, note-D4_eighth, note-D5_eighth, tie, barline,
note-D5_eighth, note-C#5_sixteenth, note-B4_sixteenth, note-C#5_sixteenth,
note-D5_sixteenth, note-E5_eighth, tie, barline, note-E5_sixteenth,
note-A4_sixteenth, note-B4_sixteenth, note-C#5_sixteenth
```

(c) Semantic encoding of the music excerpt.

Figure 1: Example of the agnostic and semantic encoding of a musical excerpt (Calvo-Zaragoza and Rizo, 2018b).

Translation Model Description

The main task of the model is to translate an agnostic sequence into its corresponding semantic sequence. We used a "seq2seq" model, first introduced by Sutskever et al. (2014) for machine translation. A seq2seq model consists of two parts, an encoder that maps the input sequence (in this case, the agnostic sequence) onto a fixed-dimension vector, and a decoder that builds the target sequence (here, the semantic sequence) from that vector. We added an attention mechanism, which has been used to improve the translation results by selectively focus on parts of the input sentence during translation (Bahdanau et al., 2014). The attention mechanism allows us to visualize which tokens (graphical symbols) of the agnostic sequence affect the translated tokens of the semantic sequence. In other words, it can show us what the model is paying attention to when translating (Figure 2).

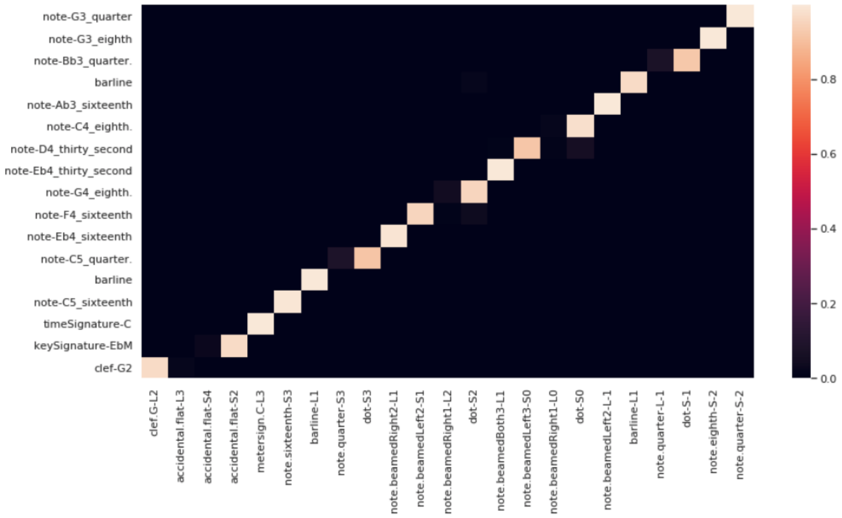
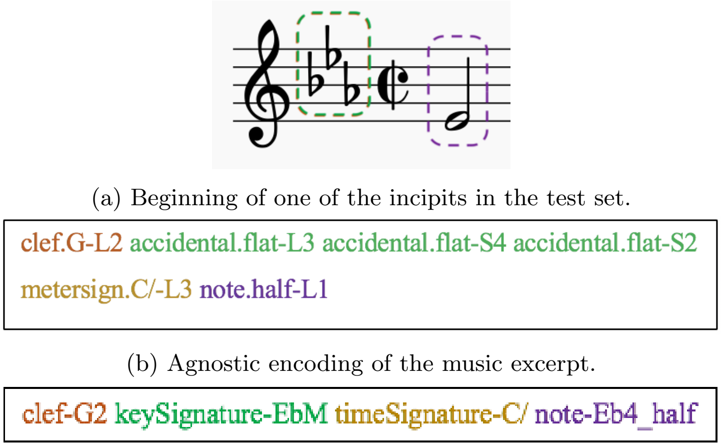


Figure 2: Attention matrix of the model when translating the agnostic sequence (horizontal axis) into a semantic sequence (vertical axis).

Experiment and Discussion

We tested this model's performance on the Printed Images of Music Staves (PrIMuS) dataset. The PrIMus dataset consists of 87,678 music incipits from RISM encoded in a variety of formats, including the agnostic and semantic representations mentioned above. We used an 80 : 10 : 10 split for training, validation, and testing.

We evaluated the model using the edit distance, which measures the number of operations (in terms of insertion, deletion, and substitution of tokens) needed for two strings to match. Given an agnostic sentence, the edit distance was computed between the corresponding semantic sequence in the dataset and the translated sequence obtained. The model flawlessly extracted the musical meaning of 85% of the agnostic sentences in the test set, correctly identifying key signatures, time signatures, multi-measure rests, dotted notes, and notes affected by notated or implied accidentals (see Figures 3 and 4).



(a) Beginning of one of the incipits in the test set.

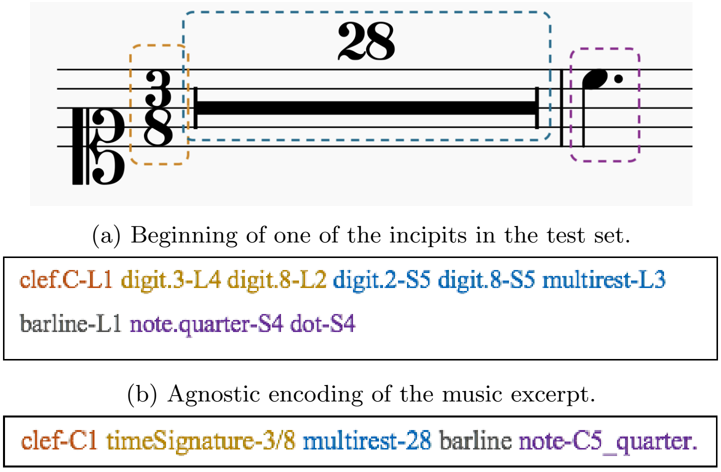
clef.G-L2 accidental.flat-L3 accidental.flat-S4 accidental.flat-S2 metersign.C/-L3 note.half-L1

(b) Agnostic encoding of the music excerpt.

clef-G2 keySignature-EbM timeSignature-C/ note-Eb4_half

(c) Semantic encoding generated by the model.

Figure 3: Example of the translation of key signatures (green) and implicit accidentals (purple) by the model.



(a) Beginning of one of the incipits in the test set.

clef.C-L1 digit.3-L4 digit.8-L2 digit.2-S5 digit.8-S5 multirest-L3 barline-L1 note.quarter-S4 dot-S4

(b) Agnostic encoding of the music excerpt.

clef-C1 timeSignature-3/8 multirest-28 barline note-C5_quarter.

(c) Semantic encoding generated by the model.

Figure 4: Example of the translation of time signatures (yellow), multi-measure rests (blue), and dotted notes (purple) by the model.

According to the edit distance values obtained, for 7% of the test sentences, only one error was made in the translation. One example of this is the sequence shown in Figure 2, where the last dotted note (coming from the agnostic tokens "note.quarter-L-1 dot-S-1") is wrongly translated into a *Bb* instead of

Ab. As seen in the attention matrix of Fig. 2, when translating dotted notes, the translator pays more attention to the dot token than to the preceding note token. Similar to dotted notes, the model also pays considerably more attention to the last accident in a series of accidentals at the moment of parsing the key signature.

As can be seen from Figure 6, most error-free sentences lie on the average-sentence-length region (the 18–33 interval with the highest data concentration in Fig. 5). Analyzing some of the examples with the highest edit distance values, some of the patterns found are the presence of a clef change, after which the translator's performance consistently drops for all following tokens; and long sentences (of more than 35 tokens, lying in the right end of the distribution shown in Fig. 5).

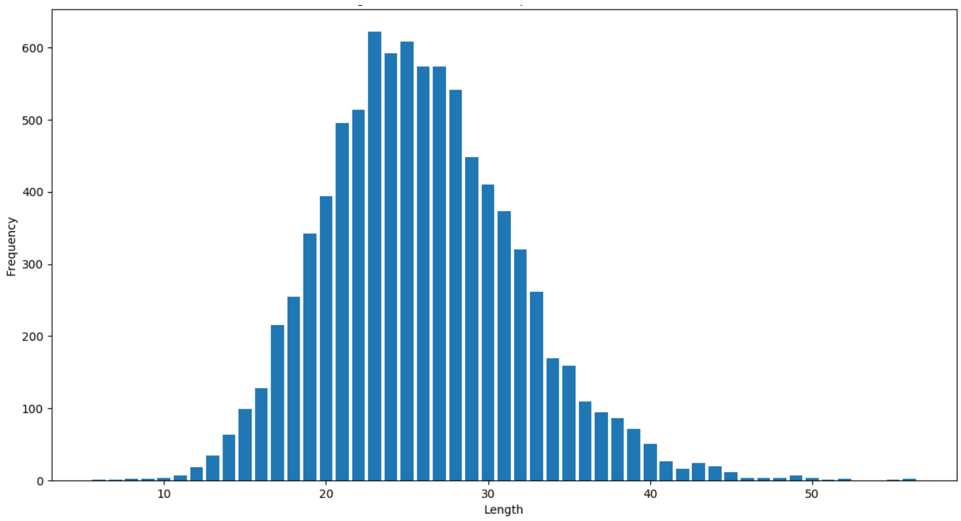


Figure 5: Length of the semantic sequences in the test set.

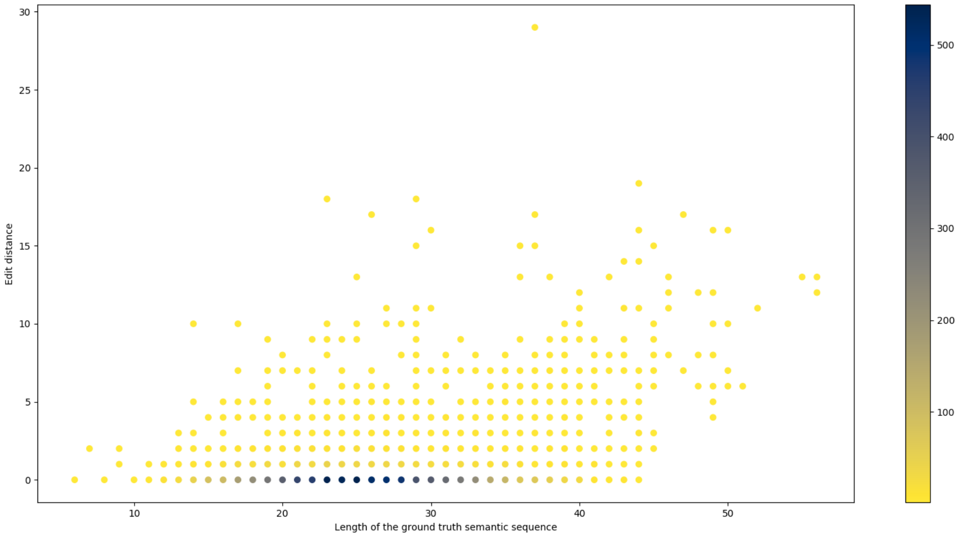


Figure 6: Color density plot of the edit distance of all sentences in the test set. The color bar indicates the frequency of a particular (sentence length, edit distance) pair.

Concluding Remarks

Given its example-based learning, the model we propose is meant to apply to different notation systems provided there is enough training data. The performance in the PrIMuS dataset was satisfactory for the vast majority of examples. However, we plan to improve the attention mechanism to enhance its performance before tackling notation systems with more complex semantics (e.g., mensural notation). Other future work includes the substitution of the semantic representation by **kern, a well-established music encoding format that also encodes the music symbols sequentially for each staff. The advantages of **kern over the semantic encoding are that the former allows for rendering the encoded sequence in Verovio, and that there is technology already available to obtain more complex formats (e.g., MEI or MusicXML) from **kern (Sapp, 2017).

Acknowledgements

This work is supported by the Spanish Ministry HISPAMUS project TIN2017-86576-R, partially funded by the EU, and by CIRMMT's Inter-Centre Research Exchange Funding and McGill's Graduate Mobility Award.

References

Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Calvo-Zaragoza, J. and D. Rizo (2018a). Camera-PrIMuS: Neural End-to-End Optical Music Recognition on Realistic Monophonic Scores. In *Proceedings of the 19th International Society of Music Information Retrieval (ISMIR)*, Paris, France, pp. 248–255.

Calvo-Zaragoza, J. and D. Rizo (2018b). End-to-End Neural Optical Music Recognition of Monophonic Scores. *Applied Sciences 8*(4), 606.

Couasnon, B. (2001). DMOS: A Generic Document Recognition Method, Application to an Automatic Generator of Musical Scores, Mathematical Formulae and Table Structures Recognition Systems. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR)*, pp. 215–220.

Rizo, D., J. Calvo-Zaragoza, J. M. Iñesta, and I. Fujinaga (2017). About Agnostic Representation of Musical Documents for Optical Music Recognition. In *Music Encoding Conference (MEC)*, Tours, France.

Rossant, F. and I. Bloch (2006). Robust and Adaptive OMR System Including Fuzzy Modeling, Fusion of Musical Rules, and Possible Error Detection. *EURASIP Journal on Advances in Signal Processing 2007*(1), 081541.

Sapp, C. S. (2017). Verovio Humdrum Viewer. In *Music Encoding Conference (MEC)*, Tours, France.

Sutskever, I., O. Vinyals, and Q. V. Le (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, pp. 3104–3112.

Szwoch, M. (2007). Guido: A Musical Score Recognition System. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, Volume 2, pp. 809–813.