The C@merata Task at MediaEval 2014: Natural Language Queries on Classical Music Scores

Richard Sutcliffe
School of CSEE
University of Essex
Colchester, UK
rsutcl@essex.ac.uk

Tim Crawford
Department of Computing
Goldsmiths, University of London
London, UK
t.crawford@gold.ac.uk

Chris Fox
School of CSEE
University of Essex
Colchester, UK
foxcj@essex.ac.uk

Deane L. Root
Department of Music
University of Pittsburgh
Pittsburgh, PA, USA
dlr@pitt.edu

Eduard Hovy
Language Technologies Institute
Carnegie-Mellon University
Pittsburgh, PA, USA
hovy@cmu.edu

ABSTRACT

This paper summarises the C@merata task in which participants built systems to answer short natural language queries about classical music scores in MusicXML. The task thus combined natural language processing with music information retrieval. Five groups from four countries submitted eight runs. The best submission scored Beat Precision 0.713 and Beat Recall 0.904.

1. INTRODUCTION

A text-based Question Answering (QA) system takes as input a short natural language query together with a document collection, and produces in response an exact answer [1]. There has been considerable progress in the development of such systems over the last ten years. At the same time, Music Information Retrieval (MIR) has been an active field for more than a decade. However, until now, there has been little or no work which draws these two fields together. The key aim of the C@merata evaluation, therefore, was to formulate a task which combines simple QA with MIR, working with Western classical art music.

In C@merata (Cl@ssical Music Extraction of Relevant Aspects by Text Analysis), participants were provided with a series of short questions referring to musical features of a corresponding score in MusicXML. The task was to identify the locations of all such features. Five groups participated. Submitted runs were evaluated automatically by reference to a gold standard prepared by the organisers.

2. APPROACH

2.1 The C@merata Task

There is a series of questions with required answers:

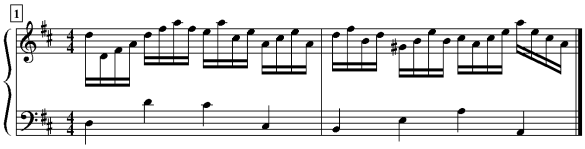
Provided **Question**:

- A short noun phrase in English referring to musical features in a score,
- A short classical music score in MusicXML.

Required **Answer**:

- The location(s) in the score of the requested musical feature.

Figure 1 shows a score extract and a corresponding question+answer. The type of the query is *followed_by*, which in this case requires G# to be followed by B. There are four answer passages, the first being `[ 4/4, 4, 2:5-2:6 ]`.



```
t: followed_by
q: G sharp followed by B
s: corelli_allegro_tr_clementi.xml
[ 4/4, 4, 2:5-2:6 ]
[ 4/4, 4, 5:15-5:16 ]
[ 4/4, 4, 8:11-8:12 ]
[ 4/4, 4, 19:1-19:2 ]
```

Figure 1. Score Extract and Example Question

The time signature is `4/4`. After this, `4` means we count in semiquavers (sixteenth notes). The passage starts in bar (measure) 2 at the fifth semiquaver and ends after the sixth semiquaver. In the task, participants are provided with the question, the score and the divisions value. They must return the answer passages. For full details of the task, see [2].

2.2 Music Scores

The music for the task was chosen from works by well-known composers active in the Renaissance and Baroque periods. The MusicXML format was chosen because it is widely used, it is relatively simple and it can capture most important aspects of a score.

For the test collection there were twenty scores, with ten questions being set for each. Scores were on one, two, three, four or five staves according to a prescribed distribution. Instrumentation was typically voices (e.g. SATB, SSA etc), Harpsichord, Lute, Violin and Harpsichord etc.

2.3 Evaluation Metrics

We adapted the well-known Precision and Recall metrics of Cyril Cleverdon which are universally used in NLP and IR. We say a passage is *Beat Correct* if it starts in the correct bar (measure) and at the right beat offset and it ends in the correct bar and at the right beat offset. Conversely a passage is *Measure Correct* if it starts in the correct bar and ends in the correct bar.

We define *Beat Precision* as the number of beat-correct passages returned by a system divided by the number of passages (correct or incorrect) returned. Similarly, *Beat Recall* is the number of beat-correct passages returned by a system divided by the total number of answer passages in the Gold Standard.

On the other hand, *Measure Precision* is the number of measure-correct passages returned by a system divided by the

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain.

number of passages (correct or incorrect) returned. *Measure Recall* is the number of measure-correct passages returned by a system divided by the total number of answer passages.

2.4 Test Queries

200 test queries were drawn up, based on twenty scores with ten questions being asked on each. American terminology (e.g. quarter note) was used for ten scores and English terminology (e.g. crotchet) for ten scores. Queries were devised in twelve different types according to a prescribed distribution as shown in Table 1 which also shows examples of each type. The Gold Standard answers were drawn up by the first author and then each file was carefully checked by one of the other authors.

Table 1. Query Types

Type	No	Example
simple_pitch	30	G5
simple_length	30	dotted quarter note
pitch_and_length	30	D# crotchet
perf_spec	10	D sharp trill
stave_spec	20	D4 in the right hand
word_spec	5	word "Se" on an A flat
followed_by	30	crotchet followed by semibreve
melodic_interval	19	melodic octave
harmonic_interval	11	harmonic major sixth
cadence_spec	5	perfect cadence
triad_spec	5	tonic triad
texture_spec	5	polyphony
All	**200**	

3. RESULTS AND DISCUSSION

3.1 Participation and Runs

Five groups from four countries (Table 2) submitted eight runs (Table 3) which were evaluated automatically using Beat Precision (BP), Beat Recall (BR), Measure Precision (MP) and Measure Recall (MR). BP and BR are much stricter, since the exact passage must be specified. However, MP and MR are also included because in practical contexts it is often sufficient to know the bar numbers - the required feature can usually be spotted very quickly by an expert.

Results were generally very good. The best was CLAS01 with Beat Precision 0.713 and Beat Recall 0.904. However, almost all runs beat the baseline run LACG01 which was prepared with the Baseline System distributed to all participants at the start [3]. Questions were intentionally easy as there were many unknown aspects of the task which had to be worked out by participants and organisers alike.

3.2 How Task was Approached

Most participants used hand crafted dictionaries and string processing to analyse the queries, rather than parsing. Generally people converted a score into feature information, extracted the required features from the query and then matched the two. Some worked with Python and Music21 while others adapted their own pre-existing systems in C++ and Common Lisp.

4. CONCLUSIONS

This was a new task at MediaEval and indeed we know of no other work combining NLP and MIR in this way. Many technical details had to be solved which sometimes took us to the limits of western classical music notation. A lot was learned from the exercise both about evaluation (e.g in devising versions of P and R to use) and about music (e.g. where does a cadence begin and end). A future task could tackle a wider range of questions involving more complicated natural language structures, as well as addressing some loose ends in the task design.

Table 2. C@merata Participants

Runtag	Leader	Affiliation	Country
CLAS	Stephen Wan	CSIRO	Australia
DMUN	Tom Collins	De Montfort University	England
OMDN	Donncha Ó Maidín	University of Limerick	Ireland
TCSL	Nikhil Kini	Tata Consultancy Services	India
UNLP	Kartik Asooja	NUI Galway	Ireland

Table 3. Results: CLAS01 is best run, LACG01 is baseline run

Run	BP	BR	MP	MR
CLAS01	**0.713**	**0.904**	**0.764**	**0.967**
DMUN01	0.372	0.712	0.409	0.784
DMUN02	0.380	0.748	0.417	0.820
DMUN03	0.440	0.868	0.462	0.910
LACG01	0.135	0.101	0.188	0.142
OMDN01	0.415	0.150	0.424	0.154
TCSL01	0.633	0.821	0.652	0.845
UNLP01	0.113	0.516	0.155	0.703
UNLP02	0.290	0.512	0.393	0.692

5. REFERENCES

[1] Sutcliffe, R., Peñas, A., Hovy, E., Forner, P., Rodrigo, A., Forascu, C., Benajiba, Y., Osenova, P. 2013. Overview of QA4MRE Main Task at CLEF 2013. *Proceedings of QA4MRE-2013*.

[2] Sutcliffe, R., Crawford, T., Hovy, E., Root, D.L. and Fox, C. 2014. Task Description v7: C@merata 14: Question Answering on Classical Music Scores. http://csee.essex.ac.uk/camerata.

[3] Sutcliffe, R. 2014. A Description of the C@merata Baseline System in Python 2.7 for Answering Natural Language Queries on MusicXML Scores. University of Essex Technical Report, 21st May, 2014.