Mind Map Using Profile Information

By Brendan Flanagan\textsuperscript{1}, Emiko Kaneko\textsuperscript{2}, Emi Izumi\textsuperscript{3},\nSachio Hirokawa\textsuperscript{4}

\textsuperscript{1} Kyushu University, JSPS Research Fellow
\textsuperscript{2} Aizu University
\textsuperscript{3} Doshisha University
\textsuperscript{4} Kyushu University
Overview

• Introduction
• Equivalent Proficiency Levels
• Grammar Pattern Item Dataset
• Mind Map Generation
• Characteristic Grammar Profiles
  • A1
  • A2
  • B1
  • B2
• Conclusion
Introduction

• **Yesterday’s presentation:** Machine learning techniques to classify the English proficiency levels of data in a spoken learner corpus.

• Machine learning techniques can classify documents into categories.

• As a part of this process, the characteristic features that are effective for classification can be extracted to create a profile of different categories.

• Example of results from machine learning:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Grammar Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.3698965</td>
<td>62-1:時制・相(過去)（be動詞）（TA.PAST.be.NEG）</td>
</tr>
<tr>
<td>2</td>
<td>-0.3412087</td>
<td>178:関係副詞(先行詞なし)（RRQ.REL.NOANT）</td>
</tr>
<tr>
<td>3</td>
<td>-0.2672319</td>
<td>193:間接話法(tell)（IND_SP.tell）</td>
</tr>
<tr>
<td>4</td>
<td>-0.2595602</td>
<td>149:that節(目的語)（CST.that.OBJ）</td>
</tr>
</tbody>
</table>
Introduction

• **Today’s presentation:** *Visualizing the relation of characteristic features of CEFR-J levels: A1, A2, B1, B2.*

• These characteristics are *not* intended for predicting differences between levels (purpose of yesterday’s talk).

• System Overview:
Equivalent Proficiency Levels

The NICT-JLE Corpus and CEFR-J

<table>
<thead>
<tr>
<th>CEFR</th>
<th>CEFR-J</th>
<th>SST</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>PreA1</td>
<td>1</td>
</tr>
<tr>
<td>A1</td>
<td>A1.1</td>
<td>2/3</td>
</tr>
<tr>
<td></td>
<td>A1.2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>A1.3</td>
<td>4</td>
</tr>
<tr>
<td>A2</td>
<td>A2.1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>A2.2</td>
<td>5</td>
</tr>
<tr>
<td>B1</td>
<td>B1.1</td>
<td>6/7</td>
</tr>
<tr>
<td></td>
<td>B1.2</td>
<td>8</td>
</tr>
<tr>
<td>B2</td>
<td>B2.1</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>B2.2</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>C2</td>
<td></td>
</tr>
</tbody>
</table>

*The NICT-JLE Corpus* is made up of 1280 transcripts of the ACTFL-ALC SST (Standard Speaking Test) English oral proficiency interview test.

There are 9 proficiency levels based on the SST scoring method.
## Equivalent Proficiency Levels

The NICT-JLE Corpus and CEFR-J

### Target Proficiency Levels:

**CEFR-J:** A1, A2, B1, B2

### SST Level 4 as CEFR-J Level A2 (in this presentation)

### Samples

<table>
<thead>
<tr>
<th>CEFR-J Level</th>
<th># Samples SST 4 as CEFR-J A1</th>
<th># Samples SST 4 as CEFR-J A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>738</td>
<td>717</td>
</tr>
<tr>
<td>A2</td>
<td>236</td>
<td>257</td>
</tr>
<tr>
<td>B1</td>
<td>263</td>
<td>263</td>
</tr>
<tr>
<td>B2</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

### Table

<table>
<thead>
<tr>
<th>CEFR</th>
<th>CEFR-J</th>
<th>SST</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>PreA1</td>
<td>1</td>
</tr>
<tr>
<td>A1</td>
<td>A1.1</td>
<td>2/3</td>
</tr>
<tr>
<td></td>
<td>A1.2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>A1.3</td>
<td>4</td>
</tr>
<tr>
<td>A2</td>
<td>A2.1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>A2.2</td>
<td>5</td>
</tr>
<tr>
<td>B1</td>
<td>B1.1</td>
<td>6/7</td>
</tr>
<tr>
<td></td>
<td>B1.2</td>
<td>8</td>
</tr>
<tr>
<td>B2</td>
<td>B2.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B2.2</td>
<td>9</td>
</tr>
</tbody>
</table>
Transcript and Tag Data

• The NICT JLE corpus exam and data structure:

<table>
<thead>
<tr>
<th>Stage</th>
<th>Task</th>
<th>Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>3</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>4</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Excluded "Follow-up" section from analysis as it contains free dialog.

• Example of tagged transcript data:

```
<text id="00086">
<stagel>
<s n="1"><w c7="UH" c5="ITJ" hw="yeah" pos="INTERJ">Yeah</w><c c5="PUN">.</c></s>
<s n="2"><w c7="APPGE" c5="DPS" hw="my" pos="PRON">my</w><w c7="NN1" c5="NN1" hw="name" pos="SUBST">name</w><w c7="VBZ" c5="VBZ" hw="be" pos="VERB">is</w><w c7="NP1" c5="NP0" hw="anonym" pos="SUBST">ANONYM</w><c c5="PUN">.</c></s>
```
Transcript and Tag Data

- The NICT JLE corpus exam and data structure:

<table>
<thead>
<tr>
<th>Stage</th>
<th>Task</th>
<th>Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>3</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>4</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Excluded "Follow-up" section from analysis as it contains free dialog.

- Example of tagged transcript data:

```
<text id="00086">
<stage1>
<s n="1"><w c7="UH" c5="ITJ" hw="yeah" pos="INTERJ">Yeah</w><c c5="PUN">.</c></s>
<s n="2"><w c7="APPGE" c5="DPS" hw="my" pos="PRON">my</w> <w c7="NN1" c5="NN1" hw="name" pos="SUBST">name</w> <w c7="VBZ" c5="VBZ" hw="be" pos="VERB">is</w> <w c7="NP1" c5="NP0" hw="anonym" pos="SUBST">ANONYM</w><c c5="PUN">.</c></s>
```

Examinee ID

CLAWS C7 & C5 tags (http://ucrel.lancs.ac.uk/claws/)

Headword

Part of Speech
Grammar Pattern Item Data

- Each section was preprocessed to count the occurrence of 493 grammar patterns, eg:

<table>
<thead>
<tr>
<th>Grammar pattern</th>
<th># 00015</th>
<th># 00253</th>
<th># 00287</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:人称代名詞主格(I)+be: I am</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1–1:人称代名詞主格(I)+be: I am not</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1–2:人称代名詞主格(I)+be: Am I ...?</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- The preprocessed tag and grammar data was then indexed to create a special purpose search engine using GETA[1].

Mind Map Generation

- Example of frequency index:

```
i24 <- document id
1 s:2 <- SST Proficiency Level
1 j:2_01139 <- Examinee id
1 p:stage1 <- SST Exam Stage
1 c:A1 <- CEFR-J Equivalent Proficiency Level
1 g:1-1 <- Grammar Pattern Item
12 o:VERB <- Part of Speech Tag
5 7:VV0 <- CLAWS C7 Tag
3 5:VVB <- CLAWS C5 Tag
1 live <- Headword
...```

Create Search Engine

Powered by GETA
Mind Map Generation

- Example of generating a mind map:

  Root

  What is Similar to the Root node?

  Search Engine
Example of generating a mind map:

Root → Node 1

Join

Best similarity

Search Engine
Mind Map Generation

• Example of generating a mind map:

Root

Node 1

What is Similar?

Search Engine

Powered by Geta
Mind Map Generation

- Example of generating a mind map:

  Repeat

```
Root
  ↓
Node 1
  ↓
Node 2
```

Best similarity

Search Engine

Powered by GETA
Mind Map Generation

- Example of generating a mind map:

```
Root

Node 1

Node 2

Node 3
```

Repeat
Mind Map Generation

• Example of generating a mind map:

Root

Node 1

Node 2

Node 3

Node 4

Repeat

Search Engine
Mind Map Generation

- Example of generating a mind map:

Repeat

Root → Node 1 → Node 2 → Node 3 → Node 4 → Node 5
CEFR-J All Levels
Overview: CEFR-J All Levels

Limits (because of size):
Words, Grammar items, C5 tags, C7 tags: <100
Overview: CEFR-J All Levels

Limits (because of size):
Words, Grammar items, C5 tags, C7 tags: <100
Grammar Profiles
Visualizing Characteristics

A Mind Map of the grammar items of CEFR-J A1

Blue outline indicates characteristic features identified in analysis by SVM

SVM Model:
CEFR-J Level A1
Grammar Profiles
Visualizing Characteristics: CEFR-J A1

A Mind Map of the word features of CEFR-J A1

Japanese words are a prominent feature of A1
Grammar Profiles
Visualizing Characteristics: CEFR-J A1

SVM Model:
Grammar Profiles
Visualizing Characteristics: CEFR-J A1
CEFR-J Level A2
Grammar Profiles
Visualizing Characteristics: CEFR-J A2

query="c:A2", 3585
Grammar Profiles
Visualizing Characteristics: CEFR-J A2

Grammar Items common across all levels

query="c:A2", 3585

SVM Model:
Grammar Profiles
Visualizing Characteristics: CEFR-J A2

SVM Model:
CEFR-J Level B1
Grammar Profiles
Visualizing Characteristics: CEFR-J B1
Grammar Profiles
Visualizing Characteristics: CEFR-J B1

query="c:B1", 1315
Grammar Profiles
Visualizing Characteristics: CEFR-J B1

SVM Model:

時制・相(現在)(一般動詞・3人称単数以外)
TA.PRESENT.do.AFF

定冠詞
AT.the

動詞-ing形
VVG

從属節(that・whether・when・if以外の主な從属接続詞)
CS.OTH
CEFR-J Level B2
Grammar Profiles
Visualizing Characteristics: CEFR-J B2

query="c:b2", 199
Grammar Profiles
Visualizing Characteristics: CEFR-J B2

No discriminative grammar item features
Grammar Profiles
Visualizing Characteristics: CEFR-J B2

SVM Model:
Grammar Profiles
Visualizing Characteristics: CEFR-J B2

SVM Model:
Conclusion

• Characteristic grammar items for each CEFR-J Level were extracted.

• In future work, we will extract the error features of spoken learner data.