Testing approaches to automatic comparison of qualifications in Poland – initial results

Thessaloniki 28.11.2019 r.

Marcin Będkowski, Wojciech Stęchły

based on joint work with:
Leopold Będkowski and Joanna Rabiega-Wiśniewska
Integrated Qualification Register

- over 10,000 qualifications (10,001 – November 28, 2019)
- ca. 700 contain full descriptions of LO’s
- ca. 500 contain short descriptions of LO’s

By the end of the year:
- several dozens of new market qualifications (over 200 in queue for next year)
- ca. 5000 descriptions for HE qualifications
- ca. 215 descriptions for VET qualifications (new curriculum)
Our Context: Qualifications Register modernization

✓ Improving searching and browsing usability
  (semantic search, filtering options, categorization and/or tagging of content, context browsing tools)

✓ Developing automatic reporting and additional queries
  (qualifications comparison, generating lists of qualifications based on selected criteria, e.g. containing phrases, similar to)

✓ Designing web applications:
  ✓ „Compass”;
  ✓ „Learning pathways”;
  ✓ „Virtual assistant”.
The „WHY?“: Similar challenges to international context?

Policy perspectives:
- Accessibility and transparency of qualifications system;
- Credit accumulation and transfer and building learning pathways (awarding bodies and learners);
- Preventing proliferation of similar qualifications;
- ...
The „WHAT?”: long list

✓ Assessing similarity of objects;
✓ Determining and representing relations between qualifications;
✓ Grouping / clustering of qualifications;
✓ Classifying and linking to existing taxonomies/classifications;
✓ Supporting decision process and qualification design/description;
✓ Supporting levelling process.
The „HOW?“: A intuitive typology of approaches (for the purpose of this presentation only)

Analytical approaches:
~ Based on separation of constituent elements of a complex entity (e.g. key features identification);
~ More formalised methodology and analysis process;
~ Conceivable output

Holistic approaches:
~ Based on analysis of whole entities;
~ Exploiting the combination of vast (yet often undefined) knowledge and heuristic reasoning;
The „HOW?”: Similar challenges?

How many people do we need to compare, group, tag 10 000 qualifications of different structure and content?

\[(X + Y) \times N\]

- X analytics
- Y qualifications experts
- for N months
Exemplary pipelines in **Orange**
Two baseline approaches

<table>
<thead>
<tr>
<th></th>
<th>Approach no. 1</th>
<th>Approach no. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basis for comparison</td>
<td>Learning outcomes</td>
<td>Synthetic description</td>
</tr>
<tr>
<td>Features</td>
<td>lemmatized nouns</td>
<td>lemmatized n-grams</td>
</tr>
<tr>
<td>No. of features</td>
<td>ca. 3300</td>
<td>ca. 4000</td>
</tr>
<tr>
<td>Feature weighting</td>
<td>0–1</td>
<td>TFIDF</td>
</tr>
<tr>
<td>Measure of similarity</td>
<td>jaccard</td>
<td>cosine</td>
</tr>
</tbody>
</table>
Natural Language Processing – basic terms

✓ lemmatization (determining base forms of words)
✓ jaccard index (of similarity)
✓ TFIDF
✓ n-grams
✓ cosine similarity
Example of data preprocessing: ‘Atomization’ of learning outcomes (LO) – difficult task for Polish

Using NLP tools, we atomized the LO’s and extracted and lemmatized relevant words:

• (The learner) *describes* and *explains* the construction of *hammers* and *nails*

  ➞

• *Describes* the construction of the *hammer* +
  *Explains* the construction of the *hammer* +
  *Describes* the construction of the *nails* +
  *Explains* the construction of the *nails*

  ➞

• *(describe, hammer, explain, nail, construction)*
Jaccard index

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \]
\[ tfidf_{i,j} = \frac{n_{i,j}}{\sum_{k} n_{k,j}} \cdot \log \frac{|D|}{|d : t_i \in d|} \]
cosine similarity

\[ \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \]
… but let us focus on the results
Case 1. Dental assistant qualification: automated extraction of keywords

Key phrases (with TFIDF values):
0.181 lekarz dentysta [dentist]
0.181 dentysta [dentist]
0.176 stomatologiczny [dental]
0.161 dentystyczny [dental]
0.152 gabinet dentystyczny [dental surgery - place]
0.150 gabinet [surgery - place]
Case 1. Dental assistant qualification: most similar qualifications based on calculation of cosine similarity

The most similar qualifications (with cosine similarity):
0.9158 – Dental hygienist
0.7508 – Assisting the dentist and keeping the surgery ready for work
0.7347 – Paramedic
0.6841 – Orthoptist
0.6777 – Dental technician
Case 2. Design of websites qualification: automated extraction of keywords

Key phrases (with TFIDF values):
0.421 server
0.386 (to) create
0.268 client
0.231 content
0.216 database
0.212 copy
Case 2. Design of websites qualification: most similar qualifications based on calculation of cosine similarity

The most similar qualifications (with cosine similarity):
0.4638 – Programming, creation and administration of websites and databases (since September 1, 2017)
0.3346 – Creation of web applications and databases and administration of databases
0.2705 – IT technician (since September 1, 2017)
0.2223 – IT technician
0.2116 – Project management
Case 3. Degglomerative (top-down) hierarchical clustering dendrogram
Case 3. Degglomerative (top-down) hierarchical clustering dendrogram – example of 4th level cluster

✓ Exploitation of mechatronic systems in agriculture
✓ Exploitation of mechatronic systems in agriculture (since September 1, 2017)
✓ Use of vehicles, machinery, equipment and tools used in agriculture
✓ Maintenance and repair of vehicles, machinery and equipment used in agriculture
✓ Maintenance and repair of vehicles, machinery and equipment used in agriculture (since 1 September 2017)
✓ Beekeeping
✓ Conducting agricultural production
✓ Running an agritourism farm
✓ Organisation and supervision of agricultural and beekeeping production
✓ Organisation and supervision of agricultural production
✓ Animal husbandry, breeding and insemination
✓ Animal husbandry and insemination (since 1 September 2017)
✓ Performing auxiliary activities in the field of veterinary services
✓ Performing auxiliary activities in the scope of veterinary inspection tasks
✓ Performing auxiliary activities in the field of veterinary services and veterinary control and supervision (since September 1, 2017)
Case 3. Deglomerative (top-down) hierarchical clustering dendrogram – example of 6th level cluster with human labelling

- Exploitation of mechatronic systems in agriculture
- Use of vehicles, machinery, equipment and tools used in agriculture
- Maintenance and repair of vehicles, machinery and equipment used in agriculture
- Maintenance and repair of vehicles, machinery and equipment used in agriculture (since September 1, 2017)
- Beekeeping
- Conducting agricultural production
- Running an agritourism farm
- Organisation and supervision of agricultural production
- Organisation and supervision of agricultural and beekeeping production
- Organisation and supervision of agricultural and beekeeping production
- Animal husbandry, breeding and insemination
- Animal husbandry and insemination (since 1 September 2017)
- Performing auxiliary activities in the field of veterinary services
- Performing auxiliary activities in the scope of veterinary inspection tasks
- Performing auxiliary activities in the field of veterinary services and veterinary control and supervision (since September 1, 2017)
Case 4. T-SNE visualisation in 2D space with K-Means clustering (colours) (demo)

https://lbedk.shinyapps.io/t-sne/
We see our results as proof of concept for:

✓ automated comparison of qualifications and explicable/interpretable degree of similarity
✓ automated extraction of key phrases for qualifications
✓ grouping / clustering of qualifications – independent from existing classifications
Work in progress

✓ grouping methods pilotage and application – testing other approaches
  ✓ knowledge-based measures using WordNet
  ✓ vector language models (word2vec, fasttext, ELMo, USE...)
  ✓ ARTM (topic modeling)
  ✓ model ensembling

✓ collecting data concerning occupations, job offers, etc. for the purpose of model training and data augmentation

✓ consultations with experts, evaluation of results

✓ feasibility study on chatbot

✓ three applications supporting register users
Thank You!

Educational Research Institute
IQS Project Office
Górczewska 8, 01-180 Warsaw, Poland
phone: +48 22 24 17 100, +48 22 24 17 111
e-mail: rejestr@ibe.edu.pl


m.bedkowski@ibe.edu.pl
w.stechly@ibe.edu.pl