Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 63-73 recibido 30-03-2023 revisado 07-05-2023 aceptado 09-05-2023

Normalisation of Education Information in Digitalised Recruitment Processes

Normalización de la Información Educativa en Procesos Digitalizados de Contratación

Laura García-Sardiña, Federico Retyk, Hermenegildo Fabregat, Lucas Alvarez Lacasa, Rus Poves, Rabih Zbib
Avature Machine Learning
machinelearning@avature.net

Abstract: Digitalised recruitment processes typically rely on key information automatically extracted from resumes. The case of educational background information is particularly noisy, considering the ever-growing naming of degrees, thus making its normalisation a decisive aspect for subsequent exploitation of such data. In this work we define the normalisation of education information as its transformation into pairs of level/field-of-study. Towards that purpose, we define and share a new taxonomy for fields of study within the labour context. We develop a simple approach where level of study is identified using expert rules, and field of study is normalised using a combination of rules to cover the most frequent occurrences and classifier predictions to generalise over the less frequent cases. We evaluate the proposed system on a new test set that we also make publicly available. We also investigate the application of education normalisation to a candidate-job matching use case.
Keywords: Education Normalisation, Parsed Resumes, Recruitment Systems, Candidate-Job Matching.

Resumen: Los procesos de contratación digitalizados suelen basarse en información clave extraída automáticamente de los currículums. El caso de la trayectoria educativa es especialmente conflictivo, considerando la creciente cantidad de titulaciones, por lo que su normalización es decisiva para la posterior explotación de dichos datos. En este trabajo definimos la normalización de la información educativa como su transformación en una serie de pares nivel-campo de estudio. Para ello, definimos y compartimos una nueva taxonomía de campos de estudio en el contexto laboral. Desarrollamos un sistema sencillo en el que el nivel de estudios se identifica mediante reglas expertas, y el campo de estudios se normaliza utilizando una combinación de reglas para cubrir las ocurrencias más frecuentes y predicciones de clasificadores para generalizar sobre los casos menos frecuentes. Evaluamos el sistema propuesto en un nuevo test set compartido públicamente y probamos su aplicación en un caso de uso de comparación candidato-empleo.
Palabras clave: Normalización de Educación, Currículums Procesados, Sistemas de Contratación, Comparación Candidato-Empleo.

1 Introduction

With the rapid digitalisation of recruitment processes, both candidates and recruiters increasingly rely on digital management systems to discover, model, and analyse relevant opportunities. Sophisticated tools such as candidate-job matching can help to optimise this process. However, the efficacy of these tools can be greatly influenced by the degree to which data are normalised. Information about the candidates is sourced through various means, including manual input, imports from external sources, and automatic extraction from their resumes. This, combined with the diversity of information expressed in natural language, implies that the input data are highly variable and noisy.

Resumes remain a key aspect of digitalised recruitment processes. They are often the first point of contact between candidates and

ISSN 1135-5948 DOI 10.26342/2023-71-5

recruiters, providing an initial impression of their suitability. Nonetheless, digitalising resumes is challenging due to their diverse formatting and structuring. The typical process for extracting information from resumes is to first convert the original file to plain text, then detect its language, and finally extract the candidate information using a language-specific resume parser. Any error introduced by a component in this pipeline adds noise to the information stored in the system.

One particularly challenging aspect of normalising resume information is *educational background*. The same degree can have different names depending on the issuing institution and country, making it difficult to compare educational backgrounds. Education programmes also evolve, with new degrees and fields of study emerging constantly. Normalisation of the education information helps to reduce this linguistic variability, which in turn improves the efficacy of down-stream tools like job-candidate matching.

In this paper, we address the task of normalising education information found in resumes by automatically mapping the diverse descriptions of education experiences to a predefined set of standard values. We formulate the problem as two short text classification sub-tasks: one for *education level*, where the target is a small set of possible classes, and one for *field of study*, where there is a large number of classes hierarchically organised. A simple approach is proposed by combining expert rules and classifier predictions.

To our knowledge, this is the first time automatic normalisation of education information is addressed in the literature. Other tasks involving the normalisation of short texts into a predefined domain-specific taxonomy include normalising products within online marketplaces (Aanen, Vandic, and Frasincar, 2015; Skinner and Kallumadi, 2019), mentions of symptoms and diseases in medical records (Jia et al., 2021; Ziletti et al., 2022), and job titles in the context of recruitment processes (Decorte et al., 2021).

This paper is organised as follows: Section 2 describes education experiences in resumes and the relevant standards for education categorisation. Section 3 describes our proposed system for normalising education information. Section 4 presents the experiments and discusses the results. Section 5 introduces an application use case, and Section 6 concludes the paper.

2 Education in Resumes

Resumes are structured into different sections, with Education being one of the most common. It lists a chronological progression of a person's educational experiences such as high school, formal academic programs, or vocational training, and it usually includes details such as the period, the institution, the education level (e.g. "bachelor degree"), the primary area of concentration (e.g. "mathematics"), and possibly other information like final grades.

In the real-world scenario that we are considering, the recruitment system follows a typical pipeline for resume digitalisation. The input resume is first converted from its original format (PDF, DOCX, DOC, RTF, etc.) into plain text. Then, a language-specific automated resume parser is used to extract the relevant information. The parser is implemented as a sequence labelling model, which classifies spans of text depending on which target entities they belong to. The parser does not perform any kind of normalisation on its own. Although the parser extracts different types of information, this work focuses on the Education section of resumes in English.

There are two main sources of variability in the digitalised resumes: the diversity in the use of language coming from the original text, and the noise introduced by the components in the pipeline. In this context, normalising these noisy data into a fixed standard becomes important for the subsequent exploitation of the extracted information.

Our main goal is to devise a system that normalises the education information extracted from resumes. For this purpose, we conduct an initial exploratory analysis on a small set of parsed data (Section 2.1) and present our definition of the normalisation task (Section 2.2).

2.1 Initial Data Exploration

In this work, we focus on the information that the resume parser identifies as `degree` (i.e. type of qualification) and `major` (i.e. area of concentration). Other extracted education-related information, like institution or period, can be considered sensitive personal information and are not necessarily required for the normalisation task.

An initial analysis was performed on 37,441 examples of (`degree`, `major`) pairs extracted from a proprietary corpus of English resumes[1]. In the observed data, `degree` shows considerately less diversity than `major`. Most of such examples refer to a small number of what can be considered canonical education degrees, such as "bachelor of arts" or "master of science". Using a reduced set of only 39 canonical degrees, around 70% of the data could be covered.

On the other hand, `major` includes highly varied text values, and is closer to an open set than to a limited set of canonical values. While some values like "economics" and "mechanical engineering" are particularly frequent, most of the samples are highly infrequent. Many factors can contribute to this: (i) the major is not frequent in the real world (e.g. "soil science"), (ii) the text is a compound of multiple majors (e.g. "business economics & public policy"), (iii) it is expressed in a non-standard way (e.g. "comp science" instead of "computer science"), or (iv) the writing contains typos (e.g. "computer sobware engineering").

An additional source of noise are errors introduced by the upstream components of the system and propagated through the rest of the pipeline. They can occur in the initial file conversion into plain text (e.g. "interna onal business") or in the resume parser (e.g. including the `degree` information as part of the `major`, "msc business intelligence").

2.2 Definition of the Normalisation Task

Several frameworks have been proposed to normalise information about education in a unified form. Notably, the International Standard Classification of Education (ISCED) was introduced by UNESCO to collect and analyse comparable education statistics, both for individual countries and internationally (UNESCO Institute for Statistics, 2012; UNESCO Institute for Statistics, 2014). ISCED is widely used in education science research and socio-economic reports (Araki, 2020; Jaunzeme and Busule, 2022).

The ISCED framework considers two classification variables: level and field of education. The **level of education** is represented as an integer scale, ranging from level 0 for early childhood education – including preschool and other forms of earlier education– to level 8 for doctorate or equivalent studies. On the other hand, the **field of education** is normalised using a 3-levels deep taxonomy with 11 broad fields, 29 narrow fields, and 80 detailed fields. Dai et al. (2015) approach a task similar to ours, by extracting LinkedIn profiles and categorising their education items into ISCED levels using keyword matching. However, they do not consider the field of education.

Our analysis of ISCED identified some issues that made it unsuitable for downstream tasks like candidate-job matching. In particular, the path distance between fields of education in this taxonomy is not a good approximation of their relationship within the labour market: "mathematics" is equally distant to "computer science" as it is to "philosophy" or "veterinary". Knowing this, alternative taxonomies were also considered for normalising the field of education. The Australian and New Zealand Standard Research Classification (ANZSRC) is a 3-level taxonomy for fields of research, with 1967 distinct fields (Hancock, 2022). The United Kingdom's Higher Education Statistics Agency proposed two taxonomies: the Joint Academic Coding System (JACS)[2], and its newer replacement, the Common Aggregation Hierarchy[3]. They are also multi-level taxonomies, with 1551 and 1092 entities respectively. The excessive fine-grained detail of these alternative taxonomies makes them unsuitable for our task as well.

For these reasons, it was decided to create a new taxonomy for fields of education, using ISCED's and others as a reference, but adapted to the labour context. The result is a 3-level taxonomy with 16 broad fields, 39 narrow fields, and 132 detailed fields. A key difference with ISCED is that detailed fields (the most specific elements) can be repeated across several groups. In this way, the path distance between fields is intended to be more informative. As an example, "biochemistry" appears both under "chemical sciences", a sibling to "chemistry", and under "biological and life sciences", a sibling to "biology".

With that in mind, we define the task of normalising an education experience as infer-

[1] No data containing personal information were used throughout the development of this work.

[2] `www.hesa.ac.uk/support/documentation/jacs`

[3] `www.hesa.ac.uk/support/documentation/hecos/cah`

ring the correct values for level and field of education based on the text coming from a resume. For level normalisation the ISCED level scale is adopted, and for field the new proposed taxonomy is used. While the preference is to map into a detailed field, which is the most specific type of element in the taxonomy, narrow and broad fields are valid normalised fields if the information observed is underspecified. Figure 1 shows an example of the input and output of education normalisation.

```
Input:
{ "degree": "bachelor of arts",
  "major": "criminal justice"  }
Output:
[ {"level": 6, "field": "law"} ]
```

Figure 1: Example of input and expected output of the education normalisation process.

3 Proposed Normalisation System

The proposed approach for the education normalisation task is described below. The normalisation of level and field of education are addressed separately, since the interdependence between these two variables is notably low.

3.1 Normalising Level of Study

The approach taken for normalising level of education for an education experience is entirely rule-based and focused on levels *3 and above*, i.e. high school education and higher, since levels below that are rarely considered in the labour market context. The logic is outlined in Algorithm 1.

Algorithm 1 Normalising level of education.

```
Input: degree, degree-to-level mappings, fallback rules
level ← ∅
degree ← clean(degree)
if degree ∈ mappings.keys then
    level ← mappings[degree]
else
    for rule in rules do
        if degree matches rule then
            level ← rule.level
return level
```

The parsed `degree` text, if any, first undergoes a basic cleaning process. This involves the normalisation of white spaces, casing and diacritics, as well as the conversion of typical abbreviations, acronyms, and variations into a standard base form[4]. For example, “b. s.”, “b.sc.”, and “bachelor of sciences” are all converted to the base form “bachelor of science”.

The resulting text is then checked against predefined degree-to-level correspondence mappings. These mappings were meticulously hand-crafted following a data-driven approach, using the exploratory `degree - major` corpus as base. If the cleaned text exists as a key in these mappings, the corresponding level of education is returned as the normalisation output. Otherwise, further generic string-matching rules are used to find the normalised level (e.g. starting with the word “vocational” or containing the string “doctor”). If no rule is matched, or if the `degree` was empty to start with, the system returns a null value.

3.2 Normalising Field of Study

The field of education is normalised using a combination of expert rules and machine learning predictions. The detailed procedure, which is described next, is shown in Algorithm 2. One thing to keep in mind is that one education experience can correspond to more that one field of study, either because of compositionality (e.g. “biology & statistics”), or because the item is interdisciplinary (e.g. “psycholinguistics”).

Algorithm 2 Normalising field of education.

```
Input: major, major-to-field mappings, classifier, degree, degree-to-field mappings
fields ← [ ]
major ← clean(major)
for part in split(major) do
    if part ∈ mappings.keys then
        field ← mappings[part]
        fields.insert(field)
    else
        pred, conf ← classifier(part)
        if conf is high then
            fields.insert(pred)
if fields is empty then
    degree ← clean(degree)
    if degree ∈ mappings.keys then
        field ← mappings[degree]
        fields.insert(field)
return fields
```

The input to this module is typically

[4]Base forms were defined during the exploratory data analysis of Section 2.1

the text extracted as `major` during parsing. The text first undergoes some pattern-based cleanup and pre-processing, which includes normalising case, spaces, connectors and punctuation (e.g. "and", "&", "/"), and diacritics. The next step is to find the fields corresponding to the clean `major` text (lines 3 to 10). If the input contains a joining connector, then the splitting function breaks it into smaller parts (e.g. "economics & management" would be split into "economics" and "management"). Each part is then checked against major-to-fields mappings, which were built in a data-driven manner and guided by the analysis of Section 2.1. If a segment cannot be mapped, it is sent to an ML-based field classifier and, only if the model's confidence is high, the prediction is accepted. The details of the classifier component are provided in Section 4.2. More information about the choice of a threshold for its integration into the full system is further explained in Section 4.3.2.

If no normalised field is found using the `major` text, then the clean `degree` text is checked against manually-defined degree-to-field mappings (lines 11-15) as a fallback. From the data exploration of Section 2.1, we know that the information parsed into the `degree` field can sometimes include relevant information about the field of education. For example, the degree "B.Eng." can suggest that the normalised field is "engineering", while the degree "master of surgery" implies "medicine" as the normalised field. Finally, if the `degree` still cannot be mapped, the normalised field is left empty.

4 Experiments

Two types of experiments are used to evaluate the effectiveness of the proposed system: testing the classifier by itself, and testing the complete system. First, alternative models for the classifier used in Algorithm 2 line 8 are examined and compared. This is the only module of the system that is trainable. Then, the overall performance of the complete system is assessed.

In particular, the following research questions are explored:

- Which family of algorithms is the most effective for the classifier component?
- Can data augmentation via synthetic data pre-training lead to performance improvements for the classifier?
- Is the information parsed as degree useful for normalising the field of education?

This section presents the datasets used for the development and evaluation of the system (Section 4.1), the experiments and results for the classifier component (Section 4.2), and those for the complete system (Section 4.3).

4.1 Datasets

The task defined in Section 2.2 is a new one with no annotated data available for either the standalone classifier evaluation or for evaluating the complete system. For this reason, we made the annotations required to conduct our experiments.

The raw data originate from a proprietary corpus of parsed resumes in English, where the `degree` and `major` information is extracted from each education experience. Three pairwise-disjoint subsets are sampled from this original set of (`degree`, `major`) pairs: $set\mathcal{A}$, $set\mathcal{B}$ and $set\mathcal{C}$.

$set\mathcal{A}$ is a large, uniformly sampled part of the original corpus. It was primarily used for exploratory data analysis (Section 2.1) and for the creation of the handcrafted mappings: degree-to-level, major-to-field, and degree-to-field. The mappings were defined in a data-driven manner to cover the most-frequent cases. For this reason, subset $set\mathcal{A}$ was not used for testing purposes.

The other two subsets are devoted to create a test set:

- $set\mathcal{B}$ was sampled uniformly from the original set. Notice that, although $set\mathcal{A}$ and $set\mathcal{B}$ are disjoint, when the `major` values are cleaned (line 2 in Algorithm 2), many of those values end up being shared across the subsets.
- $set\mathcal{C}$ is intended to be a more challenging test dataset. For this reason, it was obtained by first filtering out any pair from the original set which was fully covered by the rule-based system, and then sampling uniformly from the remaining pairs.

Gold standard annotations for $set\mathcal{B}$ and $set\mathcal{C}$ were labelled by two human annotators: each (`degree`, `major`) pair was assigned their normalised level and fields from the taxonomy. We combined the two sets to create a test set totalling 1119 education experience items. We make this test publicly available for research, together with our custom taxonomy for fields of education[5].

4.2 Evaluation of the Classifier

The classifier by itself is a machine learning model that receives a segment of text[6] and assigns it one of the possible fields from the taxonomy. We approach this as a single-label but highly multi-class classification problem.

The handcrafted major-to-field mappings are used as samples to train the classifier, as these consist of high-quality associations between clean majors and their normalised fields. We only keep one-to-one mappings and discarded the rest. We call this the **simple** training set. We also explore whether the information parsed as `degree` is useful for normalising the field of education. For that, we create a second training set called **combined**, which is also derived from the major-to-field mappings but including samples where the input is the concatenation of the clean `degree` and the clean `major`. Both these datasets are partitioned in training (90%) and development (10%) using stratified sampling.

To evaluate the classifier, the test set consisting of the joined annotations of *set*$\mathcal{B}$ and *set*$\mathcal{C}$ (see Section 4.1) is used. To make this set appropriate for testing the classifier only, we remove examples that normalise into more than one field of education, and apply the preliminary cleanup that is part of the complete pipeline. Samples found in the training set are also removed in order to evaluate only on unseen samples. Then, a **combined** version of the test set was also created to evaluate the corresponding training condition.

Finally, we explore the use of data augmentation by creating a synthetic dataset. To do so, the normalisation rules are applied on *set*$\mathcal{A}$ and only the samples that are covered by those rules are kept. These data are noisy because the application of the rules does not ensure that the output is completely correct. As an example, the text "mathematics & econoimcs" would only be assigned the field "mathematics" by the rules, since the second part of the text is not covered (due to the misspelling). We create a "simple" and "combined" version of this set, and use it to pre-train the BiLSTM model before fine-tuning on the gold training data. These conditions are labelled ftBiLSTM in the results tables.

Corpus	**Train**	**Dev**	**Test**
Simple	1789	199	488
Combined	3573	398	905
Synthetic Simple	2247	250	
Synthetic Combined	7634	849	

Table 1: Corpora generated for training and evaluating field classifiers in stand-alone mode and number of samples per partition.

Table 1 presents the corpora generated for the training and testing of the classifier component and their size.

We compare different models for the field of study classifier and report the results in Table 2. Following is a listing of the explored models. In each case, suffix -*S* refers to models trained on the simple dataset, and -*C* to those trained on the combined dataset.

SVM-S/C : Linear classifiers based on Support Vector Machines. The input features were extracted from the text and consist of the count of different char n-grams with $1 \leq n \leq 3$.

BiLSTM-S/C : Bidirectional recurrent neural network encoder based on LSTM (BiLSTM), with the input text tokenised using SentencePiece.

ftBiLSTM-S/C : BiLSTM encoder, but pre-trained using "synthetic" corpora before fine-tuning on the corresponding training set.

BERT-S/C : Transformer-based encoder, using a pre-trained BERT model.

Besides accuracy, weighted precision, recall, and F1 measure, at this level we also report on *Relaxed Accuracy*, a new metric based on the distance in the taxonomy between the predicted output and the target label. This metric stems from the intuition that not all the predictions are equally wrong when they are not an exact match. For example, predicting the field of study "history" for an item that should be normalised

[5] Available at `https://github.com/Avature/education-normalisation`

[6] As detailed in line 8 of the Algorithm 2, the input to the model is a clean segment of the major, after splitting if connectors are found.

Model	Test Simple					Test Combined				
	Rel	**Acc**	**P**	**R**	**F1**	**Rel**	**Acc**	**P**	**R**	**F1**
SVM-S	0.73	0.68	0.71	0.68	0.66	-	-	-	-	-
SVM-C	0.71	0.65	0.69	0.65	0.64	0.81	0.76	0.78	0.76	0.75
BiLSTM-S	0.57	0.49	0.60	0.49	0.50	-	-	-	-	-
BiLSTM-C	0.57	0.48	0.58	0.48	0.50	0.71	0.66	0.71	0.66	0.66
ftBiLSTM-S	0.68	0.62	0.69	0.62	0.62	-	-	-	-	-
ftBiLSTM-C	0.59	0.53	0.64	0.53	0.54	0.76	0.72	0.77	0.72	0.73
BERT-S	0.76	**0.71**	0.74	**0.71**	0.70	-	-	-	-	-
BERT-C	**0.77**	**0.71**	**0.75**	**0.71**	**0.71**	**0.86**	**0.82**	**0.84**	**0.82**	**0.81**

Table 2: Comparison of classifiers for field of education prediction on simple and combined test sets. Metrics reported include Relaxed Accuracy (Rel), Accuracy (Acc), and weighted Precision (P), Recall (R), and F1 measure. The best results are shown in bold.

as "history of art" (a sibling in the taxonomy) should not be considered as invalid as predicting "medicine" (an unrelated field). To calculate this metric, we first obtain the code[7] corresponding to the fields being compared. These codes consist of 6-character long strings which determine their position in the hierarchy: the first two characters identify the broad field, the middle two identify narrow field, and the last two the detailed field. The codes of the predicted and the true fields are compared for each hierarchy stage in a top-down manner, and the score is increased with every hierarchy level matched. The *relaxed accuracy* metric ranges from 0 –full mismatch, not even sharing the broad field level – to 1 –exact field match at all hierarchy levels–. *Relaxed Accuracy* is reported as `Rel` of Table 2.

Overall, the best results are obtained by using the BERT-based models, which are able to encode semantic information seen during the model's pre-training and fine-tuning; followed by the SVM-based models, even though the latter only consider sub-word surface form patterns. BiLSTM-based models get the worst results, although it is confirmed that pre-training models on the noisy synthetic data and then fine-tuning the model on the gold standard data helps to significantly improve the results without needing to manually annotate extra data. This is particularly true for the models trained on "simple" data, which improve by 12% in F1.

As for the research question regarding the informativeness of `degree` besides `major` information for training, Table 2 shows a general small decline in results with such an approach. The numbers are similar for SVM-based and BiLSTM-based models, with the exception of the fine-tuned BiLSTM models, for which the difference is greater. BERT-based models do get improved results when using these extended data for training, although the difference is very small.

Considering that the "combined" models learned from samples including `degree + major` texts, the performance of those models are also tested on a "combined" test set. The best results are again achieved by the BERT-based model, followed by the SVM, and fine-tuned BiLSTM.

When inspecting the wrong predictions made by the best models, the difficulty of this task becomes apparent. Many of the wrongly predicted items are samples which could be considered inter-disciplinary, and which would be hard to label by human annotators as well. For example, the BERT-S model predicted the field "geography" for the sample "geography education", while the field labelled as ground truth was "teaching". Other cases like this include "athletic medicine" –true label "medicine", predicted "sports"– or "biblical counseling" –true label "theology and religious studies", predicted "psychology and cognitive sciences"–.

4.3 Evaluation of the Complete Normalisation System

The test set described in Section 4.1 is used for the evaluation of the complete normalisation system, which includes both level and field of education.

4.3.1 Results for Level of Education

Results for the level of education normalisation are presented in Table 3. The proposed

[7] The field-to-code mappings are also shared at `https://github.com/Avature/education-normalisation`.

Acc	P	R	F1
0.96	0.97	0.96	0.96

Table 3: Accuracy (Acc) and weighted Precision (P), Recall (R), and F1 measure for level of education normalisation.

Model	Acc	P	R	F1
-	0.55	**0.91**	0.59	0.68
SVM-S	0.73	0.83	**0.80**	0.80
SVM-C	0.73	0.84	0.79	0.80
ftBiLSTM-S	0.70	0.86	0.76	0.78
ftBiLSTM-C	0.57	0.84	0.63	0.68
BERT-S	**0.75**	0.88	0.79	**0.81**
BERT-C	0.74	0.88	0.79	**0.81**

Table 4: Accuracy (Acc) and weighted Precision (P), Recall (R), and F1 measure for field of education normalisation using the complete system, comparing different models.

rule-based approach handles most of the examples found in the test set. By analysing those cases for which the system does not provide the correct answer, we find three particular situations: (i) the input `degree` contains acronyms that are not covered by the rules, e.g. "xii isc" which corresponds to Indian School Certificate; (ii) typos in the spelling of the `degree`, e.g. "hight school diploma"; (iii) and parsing errors, e.g. "engineering) c.b.s.e m.k.d.a.".

4.3.2 Results for Field of Education

Table 4 shows the results for field of education normalisation. We compare six variants of the systems using the different classification models of Section 4.2, and include a no-model condition as a baseline. When integrating the classification models into the system, an entropy threshold is used to decide when to keep or discard their predictions, as specified in Algorithm 2 line 9. For each model, a wide range of thresholds was tried over the held-out development set, finally choosing the values that achieved the best accuracy results[8].

The first observation is that using any machine learning classifier brings a notable improvement with respect to the variant that is purely rule-based. Overall, using classifiers trained on the combined dataset does not provide a significant improvement over using those trained on the simple dataset, once those are integrated into the full pipeline. The best results are obtained through the BERT-based models, closely followed by the SVM models.

Furthermore, the variants of the system using classifiers trained on the combined dataset are also evaluated in an alternative way, where the classifier receives the concatenation of the clean `degree` and the clean `major` as input (i.e. modifying the line 8 of Algorithm 2). We refer to this as the **combined pipeline**. In this way, the classifiers trained with the combined dataset are evaluated with the same type of input as they observed during training. The results are presented in Table 5. The combined pipeline does not produce significant changes in performance: the results for each variant are within 1% with respect to those of the original pipeline. The best overall results are obtained by the variant using the BERT-C model and the combined pipeline.

[8]SVM-S: 0.97; SVM-C: 0.9; ftBILSTM-S: 0.23; ftBILSTM-C: 0.2; BERT-S: 0.5; BERT-C: 0.26.

Model	Acc	P	R	F1
SVM-C	0.72	0.82	0.79	0.79
ftBiLSTM-C	0.58	0.85	0.63	0.69
BERT-C	**0.75**	0.89	0.79	**0.82**

Table 5: Accuracy (Acc) and weighted Precision (P), Recall (R), and F1 measure for field of education normalisation using the combined pipeline and models. Improvements over results presented in Table 4 in bold.

5 *Use Case: Candidate-Job Matching Based on Education*

As mentioned at the beginning of this paper, resume education normalisation is useful for several tasks. Some of these include: data analytics on a pool of candidates or on hired profiles; candidate search or filtering by education level, field, or a combination of both; or candidate-job ranking based on education. In this section we focus on the latter as an application of the proposed normalisation system in a digitalised recruitment process.

In this use case scenario, the recruitment system has a pool of job postings which include, among others, the required educational background of the candidate. We assume that these requirements are normalised into level and field of study as proposed in this paper. Using the system presented in

```
Parsed Resume:
- degree: "bachelor of business administration", major:
↪ "marketing"
- degree: "master of arts", major: "economics"
Normalised Resume:
- level: 6, field: "marketing and advertising"
- level: 7, field: "economics"

Ranked Jobs:
 1. Job title: "Business Analyst ACD". Education description: "You
 ↪ have a Master Degree in Economics, Business Administration or
 ↪ similar". Score: 1
 2. Job title: "Sales Executive". Education description: "BA/BS in
 ↪ Business, Marketing or related field". Score: 0.95
 3. Job title: "Brand Strategy Manager". Education description: "a
 ↪ bachelor's degree in Marketing or other related field".
 ↪ Score: 0.95
```

```
Job Information:
- Job title: "Sr. Scientist, Skincare Formulation, MG"
- Education description: "Bachelor degree major in Chemistry,
↪ Polymer Materials, Biology or related"

Ranked Resumes:
 1. Parsed: degree: "b.sc", major: "chemistry"
    Normalised: level: 6, field: "chemistry"
    Score: 1
 2. Parsed: degree: "bachelor's degree of sciences", major:
 ↪  "biology"
    Normalised: level: 6, field: "biology"
    Score: 1
 3. Parsed: degree: "associate's degree", major: "chemistry";
 ↪  degree: "high school diploma", major: ""
    Normalised: level: 5, field: "chemistry"; level: 3, field:
    ↪ "generic programmes"
    Score: 0.783
```

Figure 2: Examples for the tasks *rank jobs given a resume* (left) and *rank resumes given a job* (right) based on education.

this work to get the applicants' normalised education tracks as well, a simple rule-based system can be created to rank candidates for a job given its education requirements, or rank jobs for a candidate given their educational achievements. We next describe an example system with such functionality that:

- Takes into account the distance in the proposed taxonomy to score field matching, and
- Compares education items of different levels, as long as the distance is small.

Considering the second point, candidates that are under- or over-qualified with respect to the required level get a small penalisation in their scores, in order to benefit exact level matches.

Using such system, a score for each $(candidate, job)$ pair can be obtained based on their normalised education and used for ranking. Figure 2 presents some examples of the tasks for candidate-job matching based on education. Using 100 varied job postings and 180 randomly sampled resumes, the left side shows a selection of the top-ranked jobs matched for an input resume, while the right side shows the top-ranked candidates for an input job post.

The matching system used for the examples is based on expert rules, but more complex methods could be taken for developing a machine learning-based approach. A system for matching based on education could be applied in isolation for ranking tasks like the presented examples, or used as a filter to retain candidates that satisfy the minimum requirements for further selection steps. It could also be used in combination with other matching systems to create a more complex tool. For example, matching based on education, work experience, and skills could have their scores combined using different weights to get a final complex ranking that takes into account all those aspects. All these processes could help recruiters, on the one hand, by making screening processes easier with standardised information, and job-seekers on the other hand, by showing more relevant job recommendations based on their profile.

6 Conclusion

This work defined the education normalisation problem and proposed a new taxonomy for fields of study within the labour context. A simple yet effective system has been presented to approach the automation of the task of education normalisation from parsed resumes. Besides this, we have studied the effects of integrating diverse types of models –from simpler light-weight SVMs to more complex semantic-aware transformers– into the system.

Results have shown that a simple rule-based approach can account for the vast majority of cases regarding level of study normalisation, although it could miss noisier inputs like misspellings and parsing errors. With regard to field of study normalisation, a more complex task, using a combination of rules for the most frequent cases and a classifier to generalise over noisier unseen data achieves good results, even on a forcefully challenging test set.

We presented a use case where the proposed education normaliser can be integrated into recruitment management systems to help automate tasks like candidate screening or candidate-job ranking.

As far as we know, the normalisation of education-related information contained in resumes has not been addressed before. We, therefore, had to develop and validate our own datasets from a limited set of initial data in order to train and test the models. It is possible that the relatively small size of the generated corpora, and the fact that we could not sample all the classes homogeneously, have limited the scope of the conclusions reached in our results.

On the other hand, it should be noted that, given that an automatic pipeline (namely resume parsing) was used to obtain the input data, these upstream processes may have resulted in added noise to data which are already variable by nature.

Finally, the defined taxonomy for field of education and the full normalisation test set on which our proposed system was evaluated are made publicly available with this work for research purposes.

References

Aanen, S. S., D. Vandic, and F. Frasincar. 2015. Automated product taxonomy mapping in an e-commerce environment. *Expert Systems with Applications*, 42(3):1298–1313.

Araki, S. 2020. Educational expansion, skills diffusion, and the economic value of credentials and skills. *American Sociological Review*, 85(1):128–175.

Dai, K., C. G. Nespereira, A. F. Vilas, and R. P. D. Redondo. 2015. Scraping and clustering techniques for the characterization of linkedin profiles. In *CS & IT Conference Proceedings*, volume 5. CS & IT Conference Proceedings.

Decorte, J.-J., J. Van Hautte, T. Demeester, and C. Develder. 2021. JobBERT: Understanding job titles through skills. In *FEAST2021, the International Workshop on Fair, Effective And Sustainable Talent management using data science.*

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Hancock, A. 2022. Australian and new zealand standard research classification (anzsrc). *ISKO Encyclopedia of Knowledge Organization.*

Jaunzeme, I. and S. Busule. 2022. Employability of university graduates in changing labour market conditions: Case of Latvia. In *ICERI2022 Proceedings*, pages 5611–5611. IATED.

Jia, Q., D. Zhang, S. Yang, C. Xia, Y. Shi, H. Tao, C. Xu, X. Luo, Y. Ma, and Y. Xie. 2021. Traditional chinese medicine symptom normalization approach leveraging hierarchical semantic information and text matching with attention mechanism. *Journal of Biomedical Informatics*, 116:103718.

Kingma, D. P. and J. Ba. 2015. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*

Skinner, M. and S. Kallumadi. 2019. E-commerce query classification using product taxonomy mapping: A transfer learning approach. In *eCOM@ SIGIR.*

UNESCO Institute for Statistics. 2012. International standard classification of education: ISCED 2011. *Comparative Social Research*, 30.

UNESCO Institute for Statistics. 2014. ISCED Fields of Education and Training 2013 (ISCED-F 2013): manual to accompany the International Standard Classification of Education.

Ziletti, A., A. Akbik, C. Berns, T. Herold, M. Legler, and M. Viell. 2022. Medical coding with biomedical transformer ensembles and zero/few-shot learning. In *Proceedings of the 2022 NAACL-HLT: Industry Track*, pages 176–187.

A Classifiers Details

Throughout the experimentation phase, different types of models and configurations were developed and tested. Next, further details about the models reported in this work are given.

SVM-based models: Calibrated Linear SVMs. Training configuration: $C = 0.1$, $tol = 1e{-5}$, $max\_iter = 10000$.

BiLSTM-based base models: Classifiers based on BiLSTMs and using a 300-dimensional non-pretrained embedding layer. The BiLSTM generates a 300-dimensional vector space processed by a final 139-dimensional dense softmax layer. Training configuration: lr=1e−3, $dropout$=0, *Adam* optimiser (Kingma and Ba, 2015).

ftBiLSTM models: A base BiLSTM model was trained using the synthetic data. Then, the model was fine-tuned on the gold standard data. During pre-training, a learning rate of 0.01 was used, keeping the rest of the base configuration. Fine-tuning configuration: lr =1e−3, $dropout$ = 0.4, *Adam* optimiser.

BERT-based models: Classifier based on an uncased BERT (Devlin et al., 2018) architecture[9]. This model was fine-tuned on the gold-standard data using a final 139-dimensional dense softmax layer. Fine-tuning configuration: lr =3e−5, $dropout$ = 0.1, *Adam* optimiser.

[9] `https://huggingface.co/google/bert_uncased_L-12_H-768_A-12`