X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number:	761758
Project Acronym:	X5GON
Project title:	Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date:	2017-09-01 to 2020-08-31
Project Duration:	36 months
Deliverable Title:	D1.3 – Initial Content Representations
Lead beneficiary:	UCL
Type:	Report
Dissemination level:	Public
Due Date (in months):	24 (August 2019)
Date:	31-August-2019
Status (Draft/Final):	Final
Contact persons:	Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz and John Shawe-Taylor

Revision

Date	Lead author(s)	Commments
25/07/2019	Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz and John Shawe-Taylor	Initial Draft
01/08/2019	Colin de la Higuera	Added contributions from Université de Nantes
03/08/2019	Jasna Urbančič	Added the partner contributions from Institut "Jozef Stefan"
04/08/2019	Stefan Kreitmayer	Added the initial results from X5Learn Dashboard
05/08/2019	Alfons Juan	Added contributions from Universitat Politècnica de València
30/08/2019	Erik Novak	Final feedback from Institut "Jozef Stefan"
31/08/2019	Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz and John Shawe-Taylor	Final Version

Contents

1 Introduction **5**
- 1.1 Artificial Intelligence in Education . . . 5
- 1.2 Open Educational Resources . . . 5
- 1.3 X5GON . . . 6
- 1.4 Scope and Overview . . . 6

2 Related Work **7**
- 2.1 Quality Assurance of OERs . . . 7
- 2.2 Personalisation of Educational Resources . . . 7
  - 2.2.1 Features of a Good Recommendation System for Education . . . 7
- 2.3 Learning Analytics and Content Analytics . . . 8
- 2.4 Novelty in Education . . . 9
- 2.5 Learner Engagement . . . 9
- 2.6 Summary: Learner Engagement Model . . . 10

3 TrueLearn: A Bayesian Approach to Predict Engageability with OERs **12**
- 3.1 Modelling Implicit Engagement . . . 12
- 3.2 Adapting the Baselines: Modelling Knowledge . . . 12
- 3.3 TrueLearn: Introducing Novelty . . . 13
- 3.4 Processing OERs: Wikifier and the Dataset . . . 14
  - 3.4.1 Knowledge Representation . . . 14
  - 3.4.2 Lecture Transciptions and Translations . . . 15
  - 3.4.3 Final Datasets . . . 17
- 3.5 Experiments . . . 17
  - 3.5.1 Validating Wikifier . . . 17
  - 3.5.2 Experiment 1: Adapted Baselines for Modelling Background Knowledge . . . 19
  - 3.5.3 Experiment 2: Incorperating Novelty . . . 20

4 X5Learn: A Learner Facing Dashboard for Learning **21**
- 4.1 Alternative Recommendation Models . . . 21
  - 4.1.1 Probabilistic Relational Model (PRM) . . . 21
  - 4.1.2 Content Based Recommendation and Preliminary User Models . . . 22
- 4.2 System Overview . . . 22
  - 4.2.1 Content analytics . . . 22
  - 4.2.2 Learning analytics . . . 22
- 4.3 User interface . . . 23
- 4.4 Launching X5Learn . . . 23

5 Discussion and Conclusions **25**
- 5.1 Validating Content Representations . . . 25
- 5.2 TrueLearn . . . 25
- 5.3 X5Learn Dashboard . . . 26
- 5.4 Conclusions . . . 26
- 5.5 Towards Advanced Representations . . . 27

A Appendix **28**
- A.1 Problem Formulation and Assumptions for TrueLearn . . . 28

List of Figures

1 UNESCO Global Open Educational Resources Logo 5
2 X5GON 6
3 Graphical Model of the Learner Engagement Model that incorporates the drivers of learner engagement. 11
4 Factor graph for the reformulations of TrueSkill and KT and the TrueLearn model. Plates represent groups of variables. 14
5 Progress for all languages *(a)* in ASR on the left, given in terms of WER (the lower, the better) and *(b)* in SMT on the right, in terms of BLEU (the higher, the better). . 16
6 A cascading menu containing relevant KCs opens as the user hovers over a fragment in the fragments bar. The ranked tagcloud (above the fragments bar) summarises the main KCs covered in the entire resource. The information panel (below the fragments bar) contains title and metadata. 23
7 Popup view including the video, the fragments bar, description and a panel for note-taking and feedback. 24
8 Graphical representation of different assumptions that can be made when modelling learner's knowledge. The methods tested in this paper are set to test these three hypotheses. 28

List of Tables

1 Top 7 topics for the partitions of a specific lecture before weighted ranking 18
2 Top 7 topics for the partitions of the same lecture after weighted ranking 18
3 Weighted mean test performance with the 20 most active learners with top 5 topics dataset. Models labelled with (△) are trained both with positive and negative engagement labels. 19
4 Weighted mean test performance with the 20 most active learners with top 10 topics dataset. Models labelled with (△) are trained both with positive and negative engagement labels. 19
5 Mean test performance with the full dataset including all learners and models. Models labelled with (△) are trained both with positive and negative engagement labels. . . . 20

Abstract

With the emergence of Open Education Resources (OERs), educational content creation has reached a whole new scale. The quality and topical coverage of these OERs could vary significantly, affecting the user satisfaction and their engagement with these materials. This work attempts to identify different components that affect learner engagement with educational materials to utilise machine learning techniques in matching the learners to the most suitable educational resources that enhance their learning journey. Drawing inspiration from Item Response Theory and Knowledge Tracing models that are prominent among the learner modelling domains, this work extends these ideas to a more general and ambitious lifelong learning environment where learners consume diverse learning materials over longer periods of time. The background knowledge and the learning interests of the learner are inferred in union with the quality of the OERs and the novelty they introduce to the learner. Finally, these individual models that capture different drivers of learner engagement are merged together to create a lightweight, scalable online-learning learner model that can be used to predict learner engagement with OERs. The final model is also rich in interetability and can be used seamlessly with a web-based dashboard that can inform the learner about new recommendations and their own knowledge state.

1 Introduction

The global population grows in a rapid pace demanding more creative and innovative approaches to be devised in order to maintain providing high quality education to masses of learners. These learners can be diverse in many different ways including and not limited to dimensions such as cultural background, language, geographies, learning preferences and etc. Providing equal opportunities to such a diverse population can become very challenging. This has motivated the United Nations to include *Ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all*, in the sustainable development goals (SDGs) [1], the world's best plan to build a better world for people and our planet by 2030. As part of this movement, the concept *Open Educational Resources* was adopted and promoted by the UNESCO since 2002, encouraging a new family of educational resources that are geared towards democratising learning.

1.1 Artificial Intelligence in Education

In the recent years, Artificial Intelligence (AI) and Machine Learning (ML) have revolutionised how information is personalised to user needs while it shows potential in multiple domains. Among them, personalised education is very important due to the social impact it makes. AI plays a significant role in maintaining quality of education provided to learners by identifying efficient learning pathways that works best for individual learners and their learning needs.

With the recent popularity of online learning [2], we can also observe that the creation of educational resources has also increased rapidly. Promotion of online learning in both commercial (e.g. Udemy[1] and Udacity[2]) and non-commercial (e.g. Khan Academy[3] and MIT OpenCourseWare[4]) spheres have enabled the creation of an abundance of educational materials that are available in the Internet. Intuitively, large scale creation of educational material opens up opportunities for better personalisation as a wider spectrum of diverse learning resources are available.

This opportunity opens up potential to break from the traditional line of thought that heavily focuses on in-class learning to more ambitious use-cases such as Distance Learning, Massive Open Online Courses (MOOCs) and Lifelong Learning opportunities.

1.2 Open Educational Resources

Open Educational Resources can be defined as teaching, learning and research material that is available in the public domain or been published under an open license. OERs can be of any medium and the open licensing allows anyone to consume, repurpose and redistribute learning material with minimal costs and restrictions. [3].

Figure 1: UNESCO Global Open Educational Resources Logo

[1] https://www.udemy.com
[2] https://eu.udacity.com
[3] https://www.khanacademy.org
[4] https://ocw.mit.edu

The definition and scope of OERs have several working definitions. A definition that is often referred to for OERs is the William and Flora Hewlett Foundation's definition [4] which is:

> ***Open Educational Resources*** *are teaching, learning and research materials in any medium, digital or otherwise, that reside in the public domain or have been released under an open license that permits no-cost access, use, adaptation and redistribution by others with no or limited restrictions.*

This definition scopes out the domain of OERs to both digital and non-digital materials while clearly defining several types of use that are permitted within the realm of OERs. One of the prominent features of OERs is their openness, enabled through open-licensing. Open licensing in open education draws inspiration from movements in the wider context such as open knowledge, open source communities. The licensing expectations of OERs are shaped from 5R activities, namely, Retain, Reuse, Revise, Remix and Redistribute [5]. This allows generation of diverse educational resources that can target flexible, broad spectrum of learning needs that span beyond formal curricula-constrained or examination oriented learning pathways and set the stage to accommodate lifelong, informal learners.

1.3 X5GON

While popularising of OERs, another challenge that naturally has to be addressed is to identify ways to enhance adoption and ease-of-use of OERs. *X5GON: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site* ***G****lobal* ***OER*** ***N****etwork* is an initiative that attempts to develop various innovative, open technology elements that will converge the currently scattered OERs. When considering the OER ecosystem, rapid rate of creation of new resources, variable quality of resources and clustering of resources in multiple isolated silos are some of the noteworthy challenges that should be surmounted. X5GON combines various elements such as content understanding, user modelling and personalisation, and quality assurance to create a unified network of OERs across the globe. The open technologies that are anticipated to be developed through X5GON envisages access to OERs focusing on the 5 Xs, in a 1) Cross Modal, 2) Cross Cultural, 3) Cross Lingual, 4) Cross Domain, and 5) Cross Cultural setting.

Figure 2: X5GON

1.4 Scope and Overview

Through this report, we outline the work that has been carried out in the last 12 months towards building the initial content representations and modelling learners. We propose an initial content representation that is based on Wikipedia ontology. This work further attempts to build an online Bayesian algorithm that consumes the proposed content representation to predict learner engagement with OERs.

The remainder of the report goes as follows. We first investigate the related work that is relevant to developing the content representation and the learner representation in section 2. Then we proceed to section 3 where we develop the Wikipedia ontology based content representation and the Bayesian learner models that can consume these representations to produce accurate engagement predictions. In section 4, we briefly outline the alternative learner models that are being developed and proceed to outlining the interactive learning interface that is being developed in order to evaluate and utilise the recommendation models that are being developed. And Finally, we conclude our results and propose future directions in section 5.

2 Related Work

The primary goal of this work is to leverage personalised recommendation of OERs to lifelong learners. This chapter discusses the prior art that can potentially contribute towards achieving this goal. As per section 1.4, the work that relates to capturing and improving resource quality (context-agnostic) and personalised recommendations (contextual) aspects of learner-resource interaction is surveyed and discussed.

2.1 Quality Assurance of OERs

Existing work relating to using data mining and machine learning to analyse context-agnostic quality of learning resources is surprisingly scarce [6]. Majority of attempts made to formally address quality present recommendations to improve quality at the resource creation stage [7, 8].

Wikipedia utilises a review system to evaluate quality of its articles. [9] uses Support Vector Regression (SVR) to predict quality classes using text style, structure, network and review information. [10] uses similar features (and also features such as recency) but uses Ensemble methods as well which perform best in prediction task. Automatic Essay Scoring (AES) addresses a similar issue to quality assessment. Promising results have been obtained on this task through N-gram models with rank preference SVM [11] and more sophisticated deep learning models [12]. AES has a heavy focus on word tokens because topical relevance is very important in essay scoring although features such as essay length also show predictive of good essays [11]. Ranking SVM has also shown to perform in predicting readability of text [13]. Readability is also used as a raw feature for quality prediction [14].

Quality assessment of online documents go beyond educational domain and is investigated in numerous other fields [15]. Quality based ranking of documents [16], spam webpage detection [17], modelling trustability in healthcare forums [18] are several examples where we can observe quality features such as recency and textual quality resonate. Previous work has identified that Understandability, Topic Coverage, Presentation Freshness and Authority are some of the verticals quality features fall into [15].

These isolated quality features are then used to predict the engageability of educational resources as they are indexed by X5GON platform using machine learning methods. As quality of a resource affects the engageability of an educational resource, modelling quality is an essential part of building a rich learner models. A more detailed description of the work we have carried out in modelling quality is outlined in deliverables, D1.1 - Quality Assurance Models and D1.2 - Report on selected and evaluated quality assurance models.

2.2 Personalisation of Educational Resources

Recommendation systems are popular across multiple domains. Different approaches such as collaborative filtering [19], Bayesian match making[20] and extreme classification [21] are used to match resources with consumers. Contrary to conventional recommendation systems, a personalised learning system differs as there is structure/ sequence of knowledge that lead to sensible learning pathways that needs to be selected for recommendation.

2.2.1 Features of a Good Recommendation System for Education

While excelling on the personalisation front, design of a futuristic recommendation system for education should be done with additional features in mind. Different concepts are best taught using different media and modalities (text, audio, video, etc.). Lane [22] argues that a primary part of designing an effective learning resource is to choose the right media that enable the users to achieve their learning outcomes. (i) **Cross-modality** and (ii) **Cross-linguality** are vital to identifying and recommending educational resources across different modalities and languages that are most likely to help the learner. (iii) **Transparency** empowers the learners by building trust between the learner and the system while supporting the learner's metacognition processes such as planning, monitoring and reflection (e.g. Open Learner Models [23]). (iv) **Scalability** ensures that a high quality learning experience can be provided to large masses of learners over longer periods of time, essential in facilitating lifelong learning. (v) **Data efficiency** enables the system to work with less data, e.g. learning from implicit engagement data [19, 24].

2.3 Learning Analytics and Content Analytics

A personalised learning system usually consists of two main components [25]: (i) **content analytics**, which extract resource characteristics such as knowledge components (KCs) covered, quality and difficulty of the resources and (ii) **learning analytics**, which capture the learner's knowledge. In the context of learning analytics, the assessment and learning science communities focus on two paradigms: Item Response Theory [26] and Knowledge Tracing [27], which aim to assess the learner's knowledge during a limited span of time (e.g. during a test). Concerning content analytics, these insights have historically been provided by human experts. Although expert labelling appears to be a sensible solution, the rapid growth of educational resources demands for scalable automatic annotation.

Learner and resource modelling are fundamental to all adaptive educational systems. Most of the literature focuses on estimating learner's knowledge based their answers to tests [28, 29, 30]. To do so, one needs to: i) determine the skills required to solve each exercise and ii) infer the learner's knowledge state for those skills. These works model the learner at a static point in time, with a limited set of skills being assessed (in many cases, individual skills). However, for lifelong learning, a wider range of skills has to be modelled over longer spans of time and the prior research in this area is surprisingly scarce.

Content Analytics (Knowledge Components): Content representations play a key role in recommending relevant materials to learners. In an educational system, this entails extracting atomic units of learnable concepts that are contained in a learning resource. We refer to these concepts as **Knowledge Components (KCs)** that can be learned and mastered. However, KC extraction can be challenging. Expert labelling is the most commonly used approach. Although automated techniques have been proposed [31, 32], these usually rely on partial expert labelling or the use of unsupervised learning approaches [33], which are complex to tune. Advances in deep learning have also led to the proposal of deep models to learn latent KCs [34, 32] with no human knowledge engineering. However, these deep representations make the interpretability of the cognitive models and the resource representation very challenging. **Wikification**, a more recent approach, looks promising towards automatically extracting explainable KCs. Wikification identifies Wikipedia concepts present in the resource by connecting natural text to Wikipedia articles via entity linking [35]. This approach avoids expensive expert labelling while providing an ontology of humanly interpretable, domain-agnostic KCs. However, Wikipedia KCs may not be as accurate as those carefully crafted by education experts. How we use Wikification for KC extraction is detailed in section 3.4.1

Learning Analytics (Learner Skills): As per section 2.3, Learning Analytics mainly revolve around IRT and KT paradigms. IRT [26] focuses on designing, analysing and scoring ability tests by modelling both learner's knowledge and question difficulty. However, IRT does not consider changes in knowledge over time. The simplest model, known as Rasch model [26], proposes to compute the probability of scoring a correct answer as a function of the learner's skill θ_ℓ and the difficulty of the question d_r:

$$P(\texttt{correct answer}|\theta_\ell, d_r) = f(\theta_\ell - d_r), \tag{1}$$

where f is usually a logistic function. This idea has been extended to algorithms such as Elo[36], to rank chess players based on their game outcomes, where instead of having learners and resources, two players compete. Previous work has proposed the use of Elo-based algorithms for learner's modelling [37], based on its similarity to the Rasch model and its computationally light online version. The well-known TrueSkill algorithm [38] improves and extends this skill learning setting in gaming, using a Bayesian approach, allowing teams of players and adding a dynamic component to update skills over time. These ideas are directly applicable to learner knowledge assessment in a lifelong learning setting in multiple ways. 1) OERs can contain a wide range of KCs in them which will require a "team" like setting where multiple KCs and learner skills have to be modelled simultaneously, 2) necessity to model the knowledge acquisition of learners over significantly long periods of time may require adding a dynamic component, and 3) modelling populations of informal learners over long periods of time requires computationally efficient algorithms (such as online learning schemes). Through this work, we extend TrueSkill model to build **Fixed depth TrueSkill** algorithm found in section 3.2 to build a learner model that takes into account the inferred background knowledge of the learner. This model outperforms the Vanilla TrueSkill baseline model in F1-score.

KT [27] is one of the most widespread models used in intelligent tutoring systems where main difference between IRT and KT being that the difficulty of a question is not taken into account. It aims to estimate knowledge acquisition of learners as a function of practice opportunities provided through questions (a series of tests). Numerous variants of KT are emerging and showing promise, e.g. enabling individualisation [29]. More recently, Deep Knowledge Tracing [34] has used Deep Learning and shown improvement of performance over classical KT. However, the challenges in interpretability of the learned KCs can be seen as a major drawback of this approach.

2.4 Novelty in Education

Novelty is one of the important aspects that are vital to learning pathways. Contrary to a traditional recommender system that provides recommendations about similar things based on items or users, there is a trajectory that should be considered by an educational recommendation system when matching learning materials to learners. We believe that novelty experienced by the learner is one of the key drivers that can direct these trajectories. From the section 2.3 it is evident that majority of current approaches attempt to map the learner to materials based on the learner knowledge itself and content difficulty.

But when presenting with learning materials, learners intend to learn something new. Research in educational games have pointed out that learners seek satisfaction in novelty, not necessarily in increased difficulty [39]. In online chess games, it was found that players enjoyed most when there was challenge in a chess game [40]. When looking at these scenarios, novelty can be identified as an important part of learning. In section 3.3, we develop the **TrueLearn** algorithm that also takes into account novelty introduced by an OER to a learner. This model reports best results in all experiments for predicting learner engagement.

2.5 Learner Engagement

Machine-learning-based recommender systems are driven by user feedback data, such as explicit feedback from ratings or implicit feedback from user actions. In the history of recommender system research, there has been a transition from using only explicit feedback to systems that use implicit feedback of some sort [24]. Inferring user information from implicit observations is efficient for a variety of reasons. Especially in a lifelong learning situation where the learner is expected to stay with the system for a lifetime, it is highly desirable to infer user preferences from implicit observations of user interactions with the system as 1) explicit feedback is scarce and hard to collect and 2) it avoids disruptive interventions that may hinder the user experience eventually leading the user to leave the system permanently.

In educational data mining domain, several studies have shown that *learner engagement* increases the likelihood of achieving better learning outcomes both in class [41, 42] and in online learning settings [43, 44]. Engagement plays a significant role in quality of online courses as well [45, 46]. The quality of an educational resource is also indicated by its ability to enable learners to achieve better learning outcomes [22]. Due to these reasons, engagement can be used as a good proxy for high quality learner-resource interaction.

Engagement can be captured using both device-based and activity-based techniques. Activity based techniques use click streams, video view logs, etc. to heuristically measure engagement [47, 48, 44, 49]. In the context of learning and education, engagement is extensively studied in relation to learning outcomes [6, 44] and has shown that it positively attributes to it.

2.6 Summary: Learner Engagement Model

First of all, assuring quality of educational materials that are exposed to the learner is essential to the success and wide adoption of OERs. Although automating the whole quality assurance process is significantly challenging at this point, other research domains show evidence and potential of automating parts of the quality assurance process such as assessing presentation quality of educational resources.

In terms of personalising education, Content Analytics and Learning Analytics position themselves in the heart of personalised learning systems. The majority of personalised learning platforms and Intelligent Tutoring Systems have the luxury of covering a limited number of KCs crafted by domain experts due to the narrow scope they attempt to operate in (e.g. modelling learner knowledge in a specific knowledge assessment task). Contrary to that, the more ambitious lifelong learning scenario with OERs needs to rely on a broader spectrum of KCs that are inter-operable between a wide range of OERs. The research landscape also shows promise that learning schemes utilised in algorithms such as TrueSkill have the bandwidth to tackle some of the unique challenges introduced in personalising education to informal, lifelong learning (e.g. modelling multiple KCs simultaneously and incorporating a dynamic factor). These observations show that algorithms such as TrueSkill has strong potential to be reformulated and extended for lifelong learning with OERs. Personal interests and preferences has been used for information retrieval quite commonly. In education use-cases, interest can be complemented with Novelty. Novelty represents new knowledge that the learner hasn't already acquired.

From the above findings, we identify different drivers that impact the learning experience with different learning resources. Variables such as resource quality (Q), background knowledge (K) of learner, novelty (N) and curiosity (C) of the learner can explain the engagement of a learner with an educational resource. Figure 3 outlines a representation of engagement as a function of these drivers.

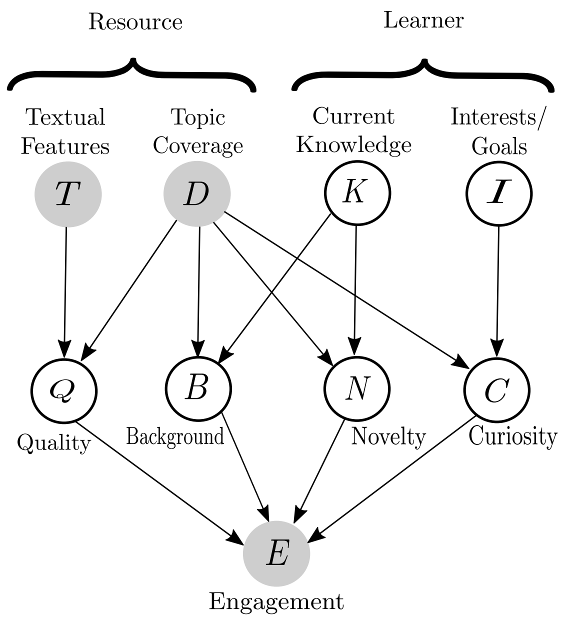

Figure 3: Graphical Model of the Learner Engagement Model that incorporates the drivers of learner engagement.

3 TrueLearn: A Bayesian Approach to Predict Engageability with OERs

One-on-one human tutoring has shown learning gains on the order of two standard deviations [50]. Recently, with the emergence of online learning platforms [2], machine learning shows promise in providing high quality personalised teaching to anyone in the world in a cost effective manner [34]. Meanwhile, Open Educational Resources (OERs) have set themselves on a fast growth trajectory, gaining popularity.

This chapter proposes **a set of Bayesian strategies aimed at providing educational recommendations to lifelong learners** using learner's engagement. We use an ontology based on Wikipedia to extract meaningful information from the text of the educational resources. Our objective is to develop an adaptive and scalable system that can recommend suitably difficult material and is transparent so that learners can check their progress. We hypothesise that there might be other factors involved in engagement apart from knowledge, such as the novelty of the material to the learner. Our approach differs from previous work in that it can be applied in situations where explicit feedback about the learner's knowledge (such as test answers) is unavailable, as tends to be the case in informal, lifelong learning. We test the different models and assumptions made using a new VideoLectures.net dataset composed of 18,933 learners and 248,643 view log entries, with promising results.

3.1 Modelling Implicit Engagement

Requiring learners to provide explicit feedback frequently can hinder the experience and discourage the use of the system. Instead, we consider the use of implicit feedback in the form of engagement. The learning algorithms presented here are aimed at predicting educational engagement.

In summary, the proposed system would recommend resources for which the learner has the necessary knowledge but there is novelty. Although other factors, such as resource quality, might play a role in engagement, we exclude these components from our model at this point, assuming that all available resources are of relatively similar quality. To be able to handle large-scale scenarios, our work focuses on online learning solutions that are massively parallelisable while prioritising on models that can be run per learner, to enhance simplicity and transparency.

3.2 Adapting the Baselines: Modelling Knowledge

Given that learning from educational engagement is a relatively novel research area, we could not find any suitable baselines to compare against. Therefore, **our first contribution is extending the two most well-known approaches for modelling skills/knowledge**: TrueSkill [38] and KT [27].

In TrueSkill, each player ℓ is assumed to have an unknown real skill $\theta_\ell^t \in R$, exhibiting a performance p_ℓ^t drawn according to $p(p_\ell^t|\theta_\ell^t) = \mathcal{N}(p_\ell^t; \theta_\ell^t, \beta^2)$ with fixed variance β^2. The outcome of the game y_{ij}^t between two players ℓ_1 and ℓ_2 (in our case learner ℓ and resource r_i) is modelled as:

$$P(p_{\ell_1}^t > p_{\ell_2}^t|\theta_{\ell_1}^t, \theta_{\ell_2}^t) := \Phi\left(\frac{p_{\ell_1}^t - p_{\ell_2}^t}{\sqrt{2}\beta}\right), \tag{2}$$

where Φ is the cumulative density of a zero-mean unit variance Gaussian. In our case, we have multiple skills associated to a learner: $\theta_\ell = (\theta_{\ell 1}, \ldots, \theta_{\ell N})$. TrueSkill allows to consider teams, assuming that the performance of a team is the sum of the individual performances of its players. For reformulating TrueSkill to our problem, we consider two approaches. The first, referred to as **Vanilla TrueSkill**, models both the skill of the learner θ_ℓ and the depth of the resources d_z as two teams playing a game, where both learners and resources are represented as a "team of skills":

$$p_\ell^t = \sum_{j \in K_{r_i}} p_{\ell j}^t, \quad p_{r_i} = \sum_{j \in K_{r_i}} p_{r_i j}. \tag{3}$$

Engagement is used as output of the game, meaning that if the learner is engaged, the skill of the learner is equal or larger than the depth of the resource $P(e_{\ell r}^t) = P(p_\ell^t > p_{r_i}^t)$. Knowledge components K_z thus define teams in our case. We consider this approach rather than assuming that each individual skill has to win over its associated KC depth because we observed that most KCs showed related topics. A similar approach using Elo system and knowledge for only one skill was considered in [37]. For the second model (named **Fixed depth TrueSkill**), we use a similar approach but fix the branch to the observed knowledge depth (using cosine similarity defined in Section 3.4).

Unlike TrueSkill, KT uses Bernoulli variables to model skills $\theta_\ell^t \sim \texttt{Bernoulli}(1, \pi_\ell^t)$, assuming that each learner ℓ would have either mastered a skill or not. Importantly, the objective of KT is not to model the learning, but capture the state of mastery of the learner at a given time, since skills are not expected to change during a test. KT further considers that once a learner has mastered a skill they cannot unlearn it. For the extension of KT (named **Multi Skill KT**), we also formulate it considering multiple skills. Skills are initialised using a $\texttt{Bernoulli}(0.5)$ prior, assuming that the latent skill is equally likely to be mastered than not. A noise factor is also included (similarly to the use of β in TrueSkill). This reformulation is inspired by the one presented in [51].

Figure 4 shows a representation of the factor graphs used for these three models, together with TrueLearn, covered in the next section. A factor graph is a bi-partite graph consisting of variable and factor nodes, shown respectively with circles and squares. Gray filled circles represent observed variables. Message passing is used for inference, where messages are approximated as well as possible through moment matching. Since our aim is to report skill estimates in real-time after learner's activity, we use an online learning scheme referred to as density filtering for all three models, where the posterior distribution is used as the prior distribution for the next time instant. The three models presented here implement the hypothesis in the central part of Figure 8, where it is assumed that we can only gather learner's knowledge from positive engagement data.

3.3 TrueLearn: Introducing Novelty

Our proposed model, **TrueLearn**, is inspired on TrueSkill with regard to representing and learning skills (we use TrueSkill at this stage because we saw in preliminary experiments that its reformulation could predict engagement better than KT). TrueLearn, additionally, introduces the aspect of novelty. Novelty is defined as the degree to which a resource contains KCs that are new to the learner. Engagement outcomes $e_{\ell r_i}^t$ between learner ℓ and resource r_i are determined in this case as:

$$e_{\ell r_i}^t := \begin{Bmatrix} +1 \ \texttt{if} |p_\ell^t - p_{\ell r_i}^t| \leq \varepsilon_\ell^t \\ -1 \ \texttt{otherwise}, \end{Bmatrix} \tag{4}$$

where the parameter $\varepsilon_\ell^t > 0$ is referred to as the engagement margin and is learner dependent. This represents the idea that both the learner and resource must be found in a similar knowledge state for the learner to be engaged (right plot in Figure 8). This engagement margin ε_ℓ^t is set counting the fraction of engaged outcomes for a learner and relating the margin to the change of engagement by:

$$P(e_{\ell r}^t) = \Phi\left(\frac{\varepsilon_\ell^t}{\sqrt{|K_r|}\beta}\right) - \Phi\left(\frac{-\varepsilon_\ell^t}{\sqrt{|K_r|}\beta}\right). \tag{5}$$

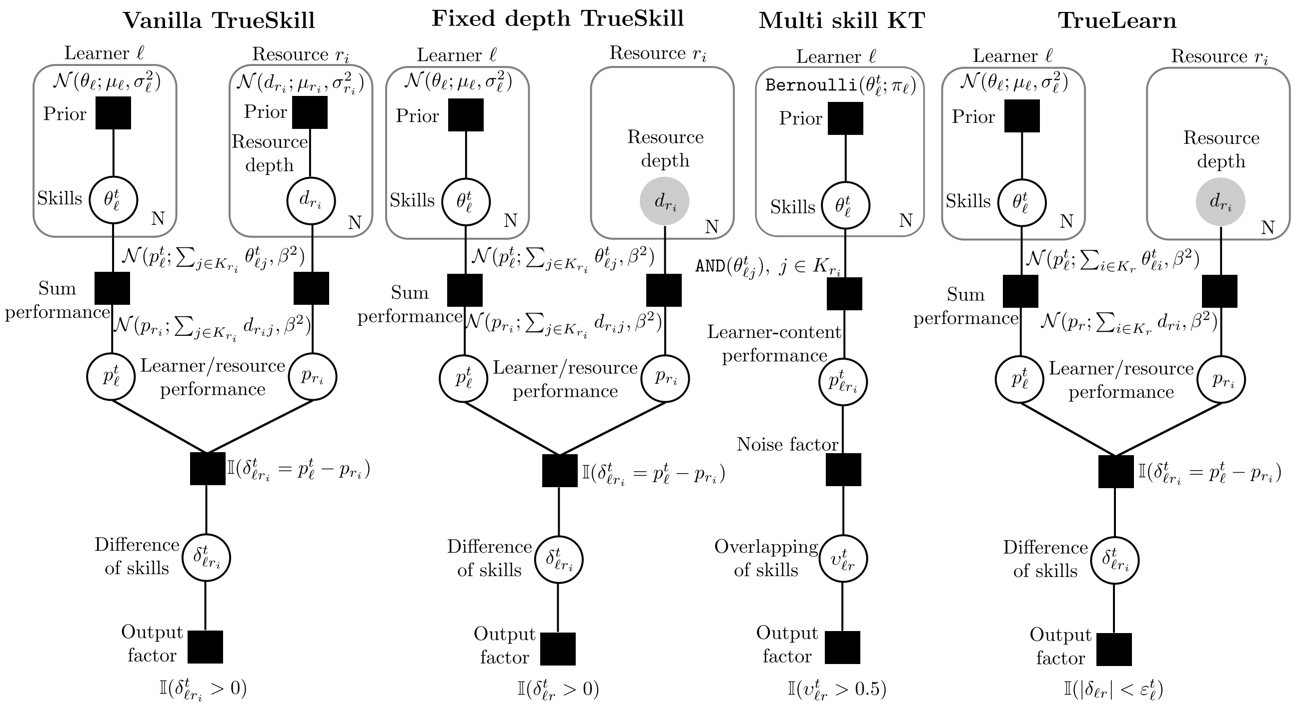

Figure 4: Factor graph for the reformulations of TrueSkill and KT and the TrueLearn model. Plates represent groups of variables.

The final model can be seen in Figure 4. This model implements the assumption in the right part of Figure 8, where we can learn learner's knowledge from both positive and negative engagement data. The function represented by the factor graphs is the joint distribution $p(\theta_\ell, p_\ell, p_r | e_{\ell r}, K_r, d_r)$, given by the product of all the functions associated with each factor. The posterior $p(\theta_\ell | e_{\ell r}, K_r, d_r)$ is computed from the joint distribution integrating the learner and resource performances p_ℓ and p_r.

Dynamics So far all models assume a stationary data distribution and hence in the limit of infinite observations, learning would come to a halt. Like in TrueSkill, we consider a Gaussian drift over skills between time steps given by $p(\theta_\ell^t | \theta_\ell^{t-1}) = \mathcal{N}(\theta^t; \theta^{t-1}, \tau^2)$. This is introduced as an additive variance component in the subsequent prior. For Multi skill KT, we increase the uncertainty by moving π_ℓ in the direction of 0.5 probability in steps of τ.

3.4 Processing OERs: Wikifier and the Dataset

We set high importance to leveraging cross-modal and cross-lingual capabilities in the desired system as these features are vital to processing all types of open educational resources in the real-world. We choose text as a generic form of raw representation for resources as the majority of modalities (videos, audio, books, web pages, etc.) can be easily converted to text. From text we extract KCs, together with the depth to which these KCs are covered.

3.4.1 Knowledge Representation

We propose to use an ontology based on Wikipedia to represent KCs. More specifically, we use **Wikifier**[5], an entity linking technique that annotates resources with relevant Wikipedia concepts [35].

[5] www.wikifier.org

Two statistics are computed for each Wikipedia topic associated with the resource: i) **PageRank score** (that represents the authority of a topic within the whole set of topics covered [52]) and ii) **Cosine similarity** (between the Wikipedia page of the topic and the resource). We use cosine similarity as a proxy for the depth of knowledge covered. Each Wikipedia topic is defined as a learnable KC. We also divide resources into what we call learnable units (fragments). A resource would then be composed of different fragments. We believe this is meaningful for two reasons: i) it enables recommending fine-grained resource fragments suited for the learner's learning path, rather than only whole resources, and ii) because in many cases the learner might not consume the resource entirely (e.g. a book), and we may want to learn exactly from the different fragments consumed.

3.4.2 Lecture Transciptions and Translations

The English translations of non-English lectures are used where the lectures are not delivered in English. The overall progress in translation achieved over the course of the project is summarised in Figure 5(a), for Automatic Speech Recognition (ASR), and Figure 5(b), for Machine Translation (MT). As can be observed in Figure 5(a), the main effort devoted in year 2 for ASR was to improve the English and Slovene ASR systems. We got consistent significant relative gains in WER for both languages. In the case of English, relative gains of 4% and 28% were achieved on the official VideoLectures.Net and poliMedia pilots, respectively. This better performance was also confirmed with additional experiments on the official IWSLT 2013 test set over which a relative gain in WER of 45% was obtained. In absolute terms, it is observed that the performance of the English system is now below the threshold of 20% absolute WER points, which is often considered a clear indication of accurate transcriptions. In the case of Slovene, on the other hand, relative improvements of 19% and 24% were achieved on, respectively, the figures reported for VideoLectures.NET (VL) and SI-TEDx-UM (TED) in year 1. It is highly remarkable that we are now much closer to the 20% WER threshold for Slovene ASR; indeed it was crossed for TED.

As in the case of ASR, from Figure 5(b) we can easily spot the language pairs we dealt with in year 2 and the evaluation results we got. Figure 5 shows the evolution of the performance of the MT systems in terms of BLEU scores (the higher, the better) for language pairs involving English (En), German (De), French (Fr) Spanish (Es), Italian (It), Slovenian (Sl) and Portuguese (Pt) on the in-domain task, VideoLectures.NET (VL) and on well-known (out-domain) tasks that are widely used for comparison purposes by the MT research community (WMT and IWSLT). Generally speaking, the focus in year 2 has been on the deployment of a series of Neural MT (NMT) systems for language pairs of special interest in the project, and also for language pairs that will be certainly needed when extending the X5GON network to sites other than those of the official pilots. Moreover, as can be observed in Figure 5(b), most systems are the first systems deployed in X5GON for their corresponding language pairs. The only exceptions are those for German-English and English-German, with relative improvements of 7% and 11%, respectively, and that for Spanish-English, with a significant 30% relative increase. In brief, many of the systems deployed exhibit BLEU scores clearly above 35, or just below 35, which is a common reference for experts to consider them good enough for practical use. For systems showing scores below 30, which includes most in-domain evaluations on VL, more effort is still required. To end this overview of MT results, we refer the reader to Deliverable D3.4 - Early support for cross-lingual OER, where comparative results with Google Translate are provided. In brief, X5GON MT systems are more or less on par with Google Translate for most language pairs, with the exception of Italian $\leftrightarrow$ English, in which Google Translate is clearly ahead of X5GON, and Slovenian $\leftrightarrow$ English and Portuguese $\leftrightarrow$ Spanish, in which X5GON MT systems clearly outperform Google Translate. To us, being far ahead of Google Translate in key language pairs such as Slovenian $\leftrightarrow$ English is a solid evidence that effective cross-lingual support for X5GON can only come from state-of-the-art MT systems adapted to the X5GON domain.

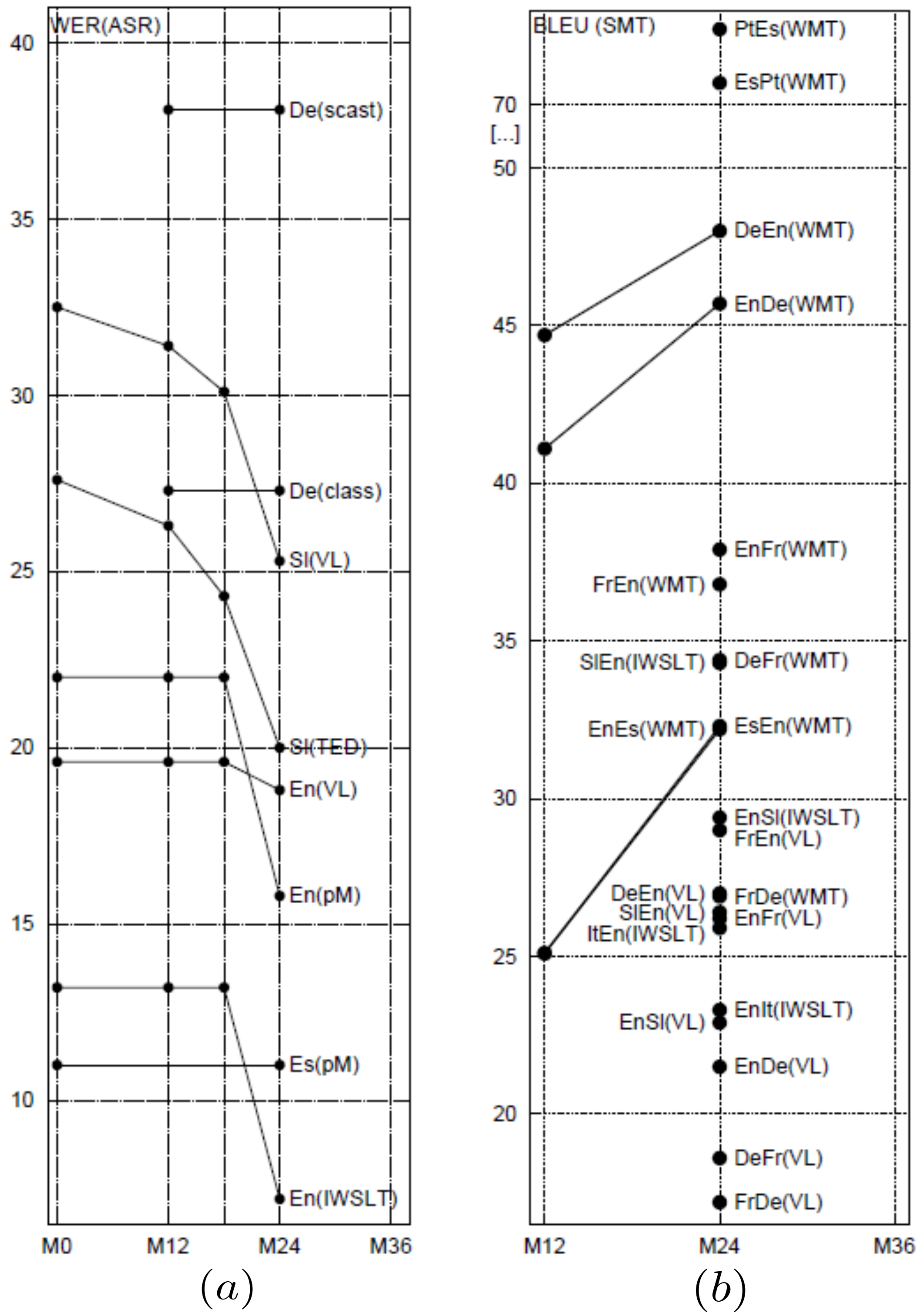

Figure 5: Progress for all languages *(a)* in ASR on the left, given in terms of WER (the lower, the better) and *(b)* in SMT on the right, in terms of BLEU (the higher, the better).

3.4.3 Final Datasets

We use a dataset which consists of users watching video lectures[6]. The lectures are also accompanied with transcriptions and multiple translations that are provided by the TransLectures project[7]. We use the English transcription of the lecture (or the English translation where the resource is non-English) to annotate the lecture with relevant KCs using Wikifier. Due to the technical limitations of Wikifier, we divide the lecture text into multiple fragments of approx. 5,000 characters. Once the lecture fragments are wikified, we rank the topics using a linear combination of pagerank and cosine similarity (further details in Section 3.5) and use the top k ranked topics (k being 5 and 10) along with the associated cosine similarity as our feature set. We define binary engagement $e_{\ell r}$ between a learner ℓ and a resource fragment r as 1 if the learner watched at least 75% of the resource fragment, and -1 otherwise. Note that user view logs are of learners actively accessing videos, i.e. when engagement is negative the learner has accessed the material but left without spending a significant amount of time on it.

The source dataset consisted of 25,697 lectures as of February 2018 that were categorised into 21 subjects, e.g. Data Science, Computer Science, Arts, Physics, etc. However, as VideoLectures.net has a heavy presence of Computer Science and Data Science lectures, we restricted the dataset to lectures categorised under Computer Science or Data Science categories only. To create the dataset, we extracted the transcripts of the videos and their viewers' view logs. A total of 402,350 view log entries were found between December 8, 2016 and February 17, 2018. These video lectures are long videos that run for 36 minutes on average and hence discuss a large number of KCs in a single lecture. The fragmentation of lectures leads to lecture fragments that are approx. 5 minutes in length.

We create three distinct datasets, based on the number of learners and top k topics selected. The first two datasets (`20 learners-10 topics` and `20 learners-5 topics`) are created using the 20 most active users and 10 and 5 topics respectively. These 20 users associate with 6,613 unique view log entries from 400 different lectures. The third dataset (`All learners-10 topics`) consists of events from all users and is composed of **248,643 view log entries distributed among 18,933 users interacting with 3,884 different lectures**. The 5 highest ranked KCs are used to represent the knowledge composition in this dataset. The dataset with 10 topics has 10,524 unique KCs while the other two datasets (20 users and all users) with top 5 ranked topics have 7,948 unique KCs.

3.5 Experiments

In this section we present the results obtained for the different datasets presented in Section 3.4.3. The experiments conducted are set to validate (among others) the use of KT against IRT inspired models (and thus the use of Gaussian and Bernoulli knowledge variables). We also validate the different components that we propose to predict engagement (knowledge and novelty) and the number of top k topics necessary to characterise a fragment using Wikipedia concepts.

3.5.1 Validating Wikifier

Analysing the results Wikifier [35] produced for several lectures we hypothesize that neither pagerank or cosine similarity alone could be used to reliably rank KCs. Pagerank seemed to be very fine-grained and prone to transcript errors. Cosine similarity, on the other hand, presented very general topics in most cases, such as 'Science', 'Data' or 'Time'. We firstly experimented with a linear combination of these two and manually validated the superior accuracy obtained (See Table 1 and 2). Such a linear combination was also proposed by the authors in [35], however they did not experience any

[6]`www.videolectures.net`
[7]`www.translectures.eu`

improvement. We then proceed to test different weights for the linear combination using our proposed version of KT (Multi skill KT) and the F1-measure. In order to find the linear weights, we executed a grid search where values between [0, 1] were assigned to the weights before training. We concluded that the best results were obtained by weighting pagerank results by 0.4 and cosine by 0.6 (cosine similarity being previously scaled to be in the same scale as pagerank).

Table 1: Top 7 topics for the partitions of a specific lecture before weighted ranking

Part.	List of first 7 topics ordered by pagerank
1	Big data, **Silicon**, Computer science, E-commerce, Mathematics, **Gene**, Silicon Valley
2	Computer science, Data science, Civil engineering, Decision theory, Terabyte, **Science fiction**, Run time (program lifecycle phase)
3	Computer science, Genome, Loss function, **Wainwright**, Probability distribution, Privacy, Mathematics
4	Loss function, Statistician, **Scrapie**, Differential privacy, Decision theory, Privacy, Power (statistics)
5	Minimax, **Scrapie**, Mathematics, Differential privacy, Saddle point, Constrained optimization, Loss function
6	Gradient descent, Convex function, Time complexity, Oracle, Algorithm, Mathematics, Convex combination
7	Differential equation, Computer science, Discrete time and continuous time, Newton (unit), Convex function, Gradient, Function space
8	Differential equation, Newton (unit), Kinetic energy, Polynomial, Geometry, Mathematics, Rate of convergence
9	Differential equation, Discretization, Stiff equation, Discrete time and continuous time, Phase space, Physics, Mathematics
10	Stochastic differential equation, Differential equation, Stochastic, **Pop music**, Mathematics, Science, Inference

Table 2: Top 7 topics for the partitions of the same lecture after weighted ranking

Part.	List of first 7 topics ordered by relevance (combination of pagerank and cosine similarity)
1	Big data, Statistics, Data science, Computer, Science, Computer science, Business
2	Statistics, Data, Data science, Science, Computer science, Scalability, Decision-making
3	Statistics, Computer science, Database, Privacy, Computer, Data, Science
4	Database, Differential privacy, Statistics, Data, Privacy, Function (mathematics), Loss function
5	Privacy, Differential privacy, Minimax, Statistics, Data analysis, Data, Mathematical optimization
6	Mathematical optimization, Gradient descent, Algorithm, Gradient, Time complexity, Function (mathematics), Convex function
7	Differential equation, Equation, Gradient, Function (mathematics), Algorithm, Discrete time and continuous time, Acceleration
8	Equation, Differential equation, Master equation, Derivative, Logarithm, Polynomial, Function (mathematics)
9	Differential equation, Equation, Momentum, Discretization, Stiff equation, Recurrence relation, Symplectic geometry
10	Differential equation, Equation, Stochastic differential equation, Control theory, Stochastic, Science, Software

Experimental design and evaluation metrics: Given that we aim to build an online system, we test the different models using a sequential experimental design, where engagement of fragment t is predicted using fragments 1 to $t-1$. Note that we both learn and predict the engagement per fragment. Since engagement is binary, predictions for each fragment can be assembled into a confusion matrix, from which we compute well-known binary classification metrics such as accuracy, precision, recall and F1-measure per individual learner. We average these metrics per learner and weight each learner according to their activity in the system. Note that most learners present an imbalanced setting, where they are mostly engaged or disengaged. Because of this, we do not use Accuracy as the main metric, but rather focus on Recall and F1. Given the large amount of algorithms tested, we use a hierarchical approach for the experiments, in which we validate a set of hypotheses for a set of more simple algorithms and then apply the conclusions extracted for our final model. For all models, each user is run separately, except for the original TrueSkill, in which we also need to model the difficulty of content and thus we require all users. Regarding initial configurations and hyperparameters, we initialised the initial mean skill of learners to 0 for all reformulations of TrueSkill. We use grid search to find the suitable hyperparameters for the initial variance while keeping β constant at 0.5. The search range for the initial variance was [0.1, 2]. For these models, initial hyper parameters are

Table 3: Weighted mean test performance with the 20 most active learners with top 5 topics dataset. Models labelled with (△) are trained both with positive and negative engagement labels.

Algorithm	Acc.	Prec.	Rec.	F1
Naïve persistence	**0.834**	**0.755**	*0.755*	**0.755**
Naïve majority	*0.823*	0.617	0.681	0.637
Vanilla TrueSkill (△)	0.695	0.641	0.726	0.669
Multi skill KT (△)	0.709	0.659	0.619	0.629
Multi skill KT	0.703	0.654	0.642	0.636
Fixed depth TrueSkill (△)	0.808	*0.704*	0.660	0.673
Fixed depth TrueSkill	0.720	0.639	**0.858**	*0.705*

Table 4: Weighted mean test performance with the 20 most active learners with top 10 topics dataset. Models labelled with (△) are trained both with positive and negative engagement labels.

Algorithm	Acc.	Prec.	Rec.	F1
Naïve persistence	**0.834**	**0.755**	*0.755*	**0.755**
Naïve majority	*0.823*	0.617	0.681	0.637
Vanilla TrueSkill (△)	0.703	0.638	0.698	0.658
Multi skill KT (△)	0.684	0.642	0.542	0.573
Multi skill KT	0.682	0.640	0.553	0.579
Fixed depth TrueSkill (△)	0.818	*0.728*	0.669	0.679
Fixed depth TrueSkill	0.709	0.634	**0.887**	*0.707*

set in the following manner. For the original TrueSkill setting (Vanilla TrueSkill), we set the same hyperparameters used in [38]. For the reformulations of KT, we run a hyperparameter grid search for the probability values of the noise factor in the range [0, 0.3]. We also tested different combinations of τ $(0.1, 0.05, 0.01)$, the hyperparameter controlling the dynamic factor. However, the results did not changed for different settings. This suggests that the dataset might still be relatively small and sparse for this factor to have an impact. The algorithms were developed in python, using MapReduce paradigm to parallelise the computation per learner.

3.5.2 Experiment 1: Adapted Baselines for Modelling Background Knowledge

We first compare the performance of the adapted versions of TrueSkill and KT proposed in Section 3.2. We compare these to two **naïve models, namely persistence and majority**. The persistence model assumes that the current state of engagement will prevail, i.e. if the learner is engaged in the current fragment, he/she will stay engaged in the future. The majority model uses majority voting to decide on engagement. We use this experiment to validate as well the necessary number of top k topics used, running the same models both for 5 and 10 topics.

Tables 3 and 4 show the results of this first experiment, where highest performance for each metric is highlighted in **bold** face and the second best in *italic*. Firstly, we can see that the naïve persistence model is very competitive. This is mainly because we are predicting fragments, and persistence has an advantage in this case. It is usually more probable that if you are engaged, you will stay engaged.

However, note that the persistence will perform trivially when recommending new resources. The algorithms labelled with △ use both positive and negative engagement labels. We run these to validate our hypothesis that no assumption can be made about negative engagement using these models (as shown in Figure 8).

As can be seen, both types of model achieve very similar performance in the case of Multi skill KT. In the case of Fixed depth TrueSkill it is better not to use negative engagement. This goes in line with our assumption. We also validate cosine similarity as a proxy for knowledge depth, as the Fixed depth TrueSkill achieves better performance than Vanilla TrueSkill, which is run for the whole dataset and infers the latent knowledge depth. The results also show very similar or improved performance when using 5 topics, which is why we use it in the subsequent experiment.

3.5.3 Experiment 2: Incorperating Novelty

After analysing the baselines for the dataset with the 20 most active learners, we evaluate now the performance using the entire dataset (18,933 users). The results can be seen in Table 5. As can be seen, TrueLearn beats all the baselines and achieves very promising performance, even better than the persistence model for recall and F1. We thus validate the necessity of considering novelty, matching the knowledge state of learners and resources. Note that in this case, TrueLearn can make use of negative engagement, given our assumption in Figure 8.

Table 5: Mean test performance with the full dataset including all learners and models. Models labelled with (△) are trained both with positive and negative engagement labels.

Algorithm	Acc.	Prec.	Rec.	F1
Naïve persistence	*0.769*	**0.631**	0.629	0.629
Naïve majority	**0.774**	0.561	*0.640*	*0.640*
Vanilla TrueSkill (△)	0.444	0.522	0.406	0.400
Multi skill KT (△)	0.501	0.491	0.194	0.257
Multi skill KT	0.500	0.490	0.197	0.259
Fixed depth TrueSkill (△)	0.657	0.581	0.401	0.459
Fixed depth TrueSkill	0.656	0.591	0.498	0.518
TrueLearn (△)	0.672	*0.608*	**0.821**	**0.677**

4 X5Learn: A Learner Facing Dashboard for Learning

Long documents, such as e-books and lecture videos and conference talks, constitute a substantial fraction among educational resources. While many of these are of high quality and potential value to learners, research in online learner behaviour has shown that long formats are often considered overwhelming and unwieldy in practice, preventing learners from engaging with these resources [53, 54]. Since engagement is a prerequisite for achieving learning outcomes [44, 41, 43], our proposed interface aims to highlight **engageable fragments** that can serve as effective entry points (or alternatives - depending on the learner's information need at hand) to the use of entire documents. The goal is to increase transparency and put the learner in control of their educational choices.

X5Learn dashboard also enables us to implement and run alternative recommendation models with real users apart from the TrueLearn algorithm developed in chapter 3.

4.1 Alternative Recommendation Models

Apart from the TrueLearn model that is described in detail in chapter 3, there are two alternative models that are being developed.

- Probabilistic Relational Model
- Content Based Recommendation and Preliminary User Models

4.1.1 Probabilistic Relational Model (PRM)

We propose to build a recommender system by using the Probabilistic Relational Model (PRM) formalism. A PRM is composed of two components: (1) a relational schema of the domain, and (2) a probabilistic model which describes the probabilistic dependencies in the domain.

Our relational schema has two entity classes called user and document, and two relationship classes, called consultation and Is-Similar-To. We also propose one dependency structure where we define the fact that one first pertinence indicator (direct pertinence) related to one document depends on the number of times this document has been consulted. Besides, we define the indirect pertinence of one document as the weighted sum of the pertinence of its similar documents (where the weight is related to the degree of similarity between both documents).

This indirect pertinence will be used to predict the interesting documents to recommend when one user is reading one target document. Right now, the relational schema and the database have been populated from the X5GON Database.

This PRM must now be implemented and tested with this database. Results will be compared with the one obtained with the recommender system actually used in X5gon project. We will also be able to learn the structure and/or the parameters of our PRM from the actual database. Our model will also be improved when new data will be available in order to take into account more interesting features from user profile.

4.1.2 Content Based Recommendation and Preliminary User Models

One of the most important services for the X5GON network is the recommendation engine developed in WP4 as it allows the users to discover and access OER material in the connected OER repositories. In year 1 we started providing content based recommendations and extended the service to support personalized recommendations based on user models in year 2.

We provide content based recommendations using k nearest neighbors algorithm on either the text (content) or Wikipedia concepts found in the material depending on the type of query. We opted for content based recommendations in the first year as this has allowed us to collect user activity data to develop user models and thus provide personalized recommendations.

Currently we use two approaches for personalization: user-material similarity and collaborative filtering. With user-material similarity we first embed user into the same semantic space of Wikipedia concepts as the materials based on user's viewing history, and then do k nearest neighbors to find relevant materials. In collaborative filtering we search for users that have seen the same materials as the user and base the recommendations on their viewing history. Both of this approaches are appropriate for real-time updates based on user activity, however both suffer from the so called cold start problem when a new user appears – in that case we use content-based recommendations.

We explain the details of all approaches currently in use, the data we use to make recommendations, and our plans for evaluation in the following deliverables:

- D4.1 – Early prototype of user modelling architecture
- D4.2 – Final prototype of user modelling architecture
- D4.3 – Early prototype of recommendation engine
- D4.4 – Final prototype of recommendation engine

4.2 System Overview

In order to recommend relevant fragments to learners, our solution leverages **content analytics** to extract characteristics from resources, such as Knowledge Components (KCs) covered [27] and various metrics of quality and difficulty. Moreover, **learning analytics** are applied to capture the user's knowledge state and growth over time.

In this section, we describe the proposed system regarding content analytics, learning analytics and the user interface.

4.2.1 Content analytics

In a pre-processing step, the system ingests the text representation of educational resources (e.g. video transcripts) and partitions them into fragments of approximately 5 minutes. Each fragment is then annotated using **Wikifier**[8]. This approach is domain-agnostic, avoids the need for expensive expert labelling and results in human-interpretable annotations that we use as KCs.

4.2.2 Learning analytics

Any recommendation algorithm that provides probabilistic predictions can be incorporated with the proposed user interface. For the purpose of this demonstration, we use TrueLearn (See section 3.3), a probabilistic algorithm that recommends educational resources to lifelong learners, using engagement signals to build a dynamic learner model.

[8] www.wikifier.org

TrueLearn was validated using a large and new VideoLectures.Net dataset composed of 18,933 learners and 248,643 video view entries, showing very promising results when compared to related baselines. While the current model learns only with implicit data in the form of engagement, it can also be easily extended to consider explicit feedback, such as ”too difficult” or ”too easy”.

4.3 User interface

The user interface aims to augment and extend how users can engage with detailed content recommendations. A primary goal was to make recommendations transparent, informative, enjoyable to use, specific and time saving. In order to make the interface intuitive to use, our design leverages familiar patterns and techniques, such as cards, popups, and cascading menus. In addition, we introduce two novel elements, a **ranked tagcloud** and a **fragments bar**, in order to enable the learner to quickly preview KCs, as illustrated in Figure 6.

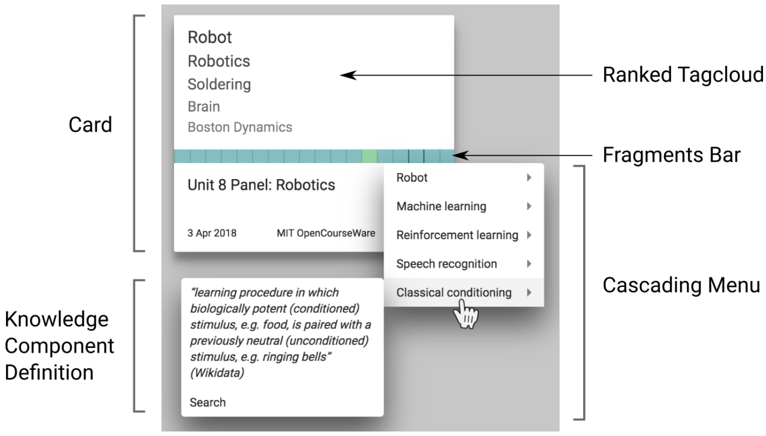

Figure 6: A cascading menu containing relevant KCs opens as the user hovers over a fragment in the fragments bar. The ranked tagcloud (above the fragments bar) summarises the main KCs covered in the entire resource. The information panel (below the fragments bar) contains title and metadata.

Selecting a fragment opens a detailed view with the video playing from the corresponding time as shown in Figure 7. In this view, the user can take notes and provide explicit feedback, e.g. ”too hard” or ”too easy”. The fragments bar allows fluid preview, recap and navigation within the resource. Different colour intensities (yellow) indicate the predicted relevance of each fragment to the learner.

Iterative design and evaluation with real users led to insights into learners’ expectations and preferences. One key finding was that staying in context is important for learners. Therefore, a pop up was used rather than page redirection.

4.4 Launching X5Learn

In order to investigate how learners can use and benefit from OER fragments and fine-grained Wikipedia annotations, user studies have been conducted at key stages of the design process, using the X5Learn dashboard with lifelong learners in the lab and in the wild.

Lab-based evaluations have primarily relied on qualitative methods, including observation-based usability evaluation, user journeys, cognitive walkthroughs and interviews. In addition, several versions of the live X5Learn website have been tested in the wild using remote user feedback. A more detailed description of these experiments is presented in deliverable D6.2.

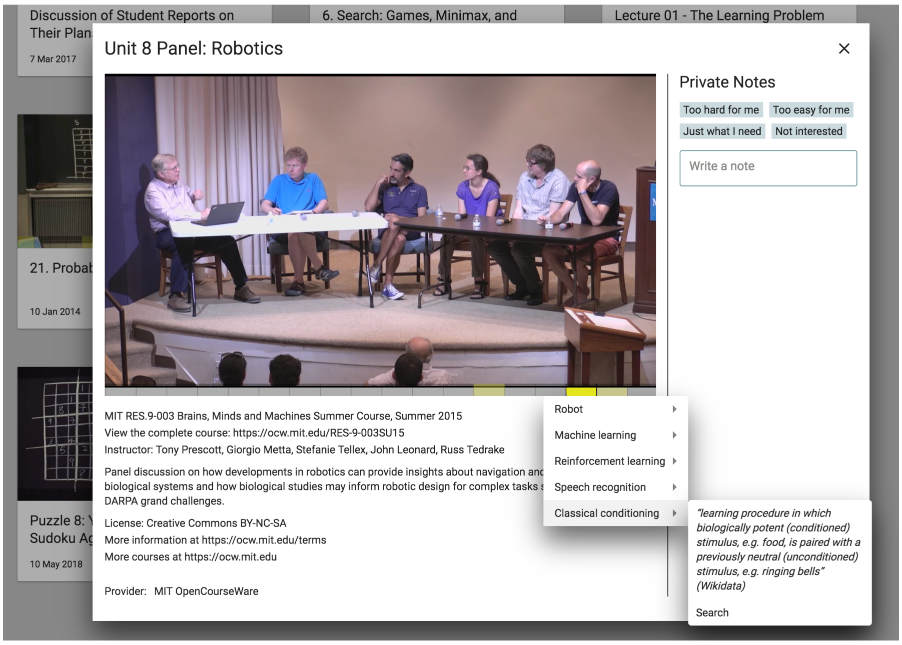

Figure 7: Popup view including the video, the fragments bar, description and a panel for note-taking and feedback.

In the near future, visitors from the general public will be encouraged to explore learning resources using a search field in addition to recommendation features. This is possible because the interface fully generalises to information search tasks. The website will be open for the public to sign up and use for lifelong learning.

5 Discussion and Conclusions

In this section we discuss the content representations and learner models that we have developed to this point.

5.1 Validating Content Representations

There are two main aspects of content representations that requires discussion. Namely, (1) Fragmentation of video lectures before annotation and (2) Wikification based content annotation itself.

Fragmentation of educational resources that opens up multiple opportunities in terms of learner modelling.

- Finer-grained visibility of user engagement
- Detailed annotations leading to richer learner models

By fragmenting a long lecture into multiple smaller fragments, we can update the learner model more frequently and in higher detail as multiple engagement signals that come from individual fragments of a long video can be used to update the model. The fragmentation also enables richer annotation of the content. Table 2 shows further evidence about the expressiveness when it comes to encoding what knowledge that is contained in individual video fragments. It is also seen how the knowledge components evolve within a long video. Table 2 clearly shows how the lecture starts with very generic concepts such as "Big Data", "Computer Science" and "Business" and over consequent fragments, evolve into finer grained knowledge components such as "Differential Privacy", "Gradient", "Derivative". The fragmentation also allows creating video fragments that are on average 5 minutes that shown evidence in video lecture literature to be significantly more engaging compared to long videos that span for 30-60 minutes [53].

The results in section 3 shows strong evidence that the content representations based on Wikipedia ontology is effective in learner modelling. Table 2 also shows the richness human interpretable nature of these representations. Furthermore, this content representation is ideal for the lifelong learning setting as the Knowledge in Wikipedia also evolves over time. In other words, our content representation choice is robust to the ever evolving nature of knowledge as the content representations also evolve over time with Wikipedia. Table 1 and 2 shows the importance of re-ranking Wikipedia concepts. Although the authors only had to consider the PageRank score in ranking topic in order to get optimal results for the disambiguation task they were solving [35]. But, our experiments show that accounting for cosine similarity between the educational resource and the Wikipedia topic page also plays a vital role in re-ranking the Wikipedia topics appropriately for the learner modelling task.

5.2 TrueLearn

In the TrueLearn framework outlined in section 3, recommendation algorithms need to focus on making recommendations for which i) the learner has enough background knowledge so they are able to understand and learn from the recommended material, and ii) the material has enough novelty that would help the learner to improve their knowledge about the subject.

Our results using a very large dataset show the potential of such an approach and its promising results. TrueLearn algorithms demonstrate the ability to outperform all its ancestors and the baselines in Recall and F1 while having minor reductions in precision. TrueLearn also embeds scalability, transparency and data efficiency in the core of its design showing clear promise towards building an effective lifelong learning recommendation system. The fact that the design of TrueLearn also has features such as scalability, transparency and data efficiency shows clear promise towards building a light online algorithm that can leverage lifelong learning at scale.

5.3 X5Learn Dashboard

The ability to use fragments and annotations in order to search for information and preview documents efficiently was generally met with great enthusiasm. All users were immediately able to use the fragments bar and cascading menus. A frequently observed pattern was that the fragments bar was used to ”speed read” through documents in order to assess their relevance. For the purpose of making informed choices, the Wikipedia annotations were generally found easy to use and occasionally more informative than human-generated titles and summaries. All test users particularly appreciated the ability to jump to specific fragments within a lecture that they found interesting.

5.4 Conclusions

While there has been vast amount of work in the context of recommendation, education domain has unique challenges that are not addressed in conventional recommendation approaches. Due to this reason, most existing algorithms tend not to be directly applicable to OER recommendation for lifelong learning. This makes TrueLearn the stepping stone to a family of data-efficient, transparent, online learner models that can be utilised to provide recommendations at scale.

The work outlined in section 3 sets the foundations towards building a lifelong learning recommendation system for educational resources. We present a empirically tested method for using Wikipedia ontology based Knowledge Components to represent educational resources. Empirical results indicate that using a linear weighting of Page Rank score and Cosine Similarity is ideal for re-ranking the Wikipedia concepts for the task of predicting learner engagement with educational resources. The presented content representation is human interpretable, domain agnostic and show to be performant with personalised recommendation algorithms. The Wikifier based content representation is also robust and automatically adapts to the evolving nature of knowledge itself. We present three novel approaches, inspired by Item Response Theory and Knowledge Tracing for making individualised prediction of learner engagement with OERs. Our proposed model (TrueLearn) introduces the concept of novelty as a function of learner engagement and show evidence of outperforming the baselines in a OER video lectures dataset in predicting learner engagement of 18,933 individual learning journeys.

While developing novel algorithms to manage lifelong scenarios is helpful, there is also a need for ambitious interfaces that can transform how we think about lifelong learning. Our design described in section 4 demonstrates a promising step in this direction by emphasising personalisation and redefining the atomic unit of educational content. It reconceptualises large educational materials as collection of building blocks that can be partitioned, recombined and re purposed effectively in non-traditional learning situations. New research questions are provoked regarding content representations and user modelling, since more fine-grained content representations enable more specific recommendations and richer engagement signals that can (and should) be incorporated into dynamic user models.

5.5 Towards Advanced Representations

Future design and research towards advanced representations should address a series of topical challenges.

Improving Content Representations We plan to investigate the functionality of Wikifier [35] thoroughly to identify what approaches can enable more robust extraction of KCs. Furthermore, we aim to incorporate the Wikipedia graph and category hierarchy to mine dependencies between KCs. We believe that incorporation of topical relationships will significantly improve modelling knowledge state, interests and novelty aspects of learners with respect to resources.

Including Learner Interests Interests and goals of the learner has a huge influence on their preference to engage with future educational resources. Prior work has also shown connection between learner interests and the openness to different degrees of novelty [55]. Therefore, it is essential that we include interest modelling to evolve the existing models into advanced learner models.

Richer Topic Coverage Features in Quality Modelling The topic coverage features utilised in deliverables - D1.1 and D1.2 (e.g. document length and document entropy) avoid deep analysis of the topic structures within an educational resource. We aim to leverage Wikifier and Wikipedia graph to build more sophisticated topic coverage features that are likely to improve quality modelling.

Launching X5Learn dashboard to the general public Simultaneously, we plan to launch the X5Learn dashboard to the general public. This will enable us to engage with real users, the ”true learners” we are trying to serve.

The advantages of deployment of the X5Learn dashboard are bi-directional. It benefits both the researchers and the learners in numerous ways.

To the Researchers:

- Providing a platform to test the TrueLearn models in the wild
- Providing a platform to compare and contrast the learner models
- Gives more control over collecting better implicit and explicit feedback
- Capacity to focus on the full solution than building parts of it in isolation

To the learners (general public):

- Provides users with learning support tools such as bookmarking, note taking capabilities
- Provides a more interactive interface for the learners to engage with the broader X5GON project and its tools

A Appendix

A.1 Problem Formulation and Assumptions for TrueLearn

Consider a learning environment in which a learner ℓ interacts with a set of educational resources $S_\ell \subset \{r_1, \ldots, r_Q\}$ over a period of $T = (1, \ldots, t)$ time steps, Q being the total of resources in the system. A resource r_i is characterised by a set of top KCs or topics $K_{r_i} \subset \{1, \ldots, N\}$ (N being the total of KCs considered by the system) and the depth of coverage d_{r_i} of those. The key idea is to model the probability of engagement $e_\ell^{r_i} \in \{1, -1\}$ between learner ℓ and resource r_i at time t as a function of the learner skill θ_ℓ^t and resource representation d_{r_i} for the top KCs covered K_{r_i}. According to Bayes rule the posterior distribution is proportional to:

$$P(\theta_\ell | e_\ell, S_\ell, K_r, d_r) \propto P(e_\ell | \theta_\ell, S_\ell, K_r, d_r) \cdot P(\theta_\ell). \tag{6}$$

Figure 8 shows the intuition behind different assumptions that can be made when modelling learner's skills. The left plot shows the assumption made in IRT and KT (both focused on test scoring rather than engagement prediction). This is, if the learner answers correctly to a test, the skill must exceed the difficulty of the question. The middle plot shows engagement as a function of knowledge, in which we hypothesise that if the learner is engaged, they have enough background knowledge to make use of the resource. However, no assumption can be made from the non-engaged cases. The last plot shows the combination of knowledge and novelty: if the learner is engaged, they must have the appropriate background to use the resource and the content must also be novel to them.

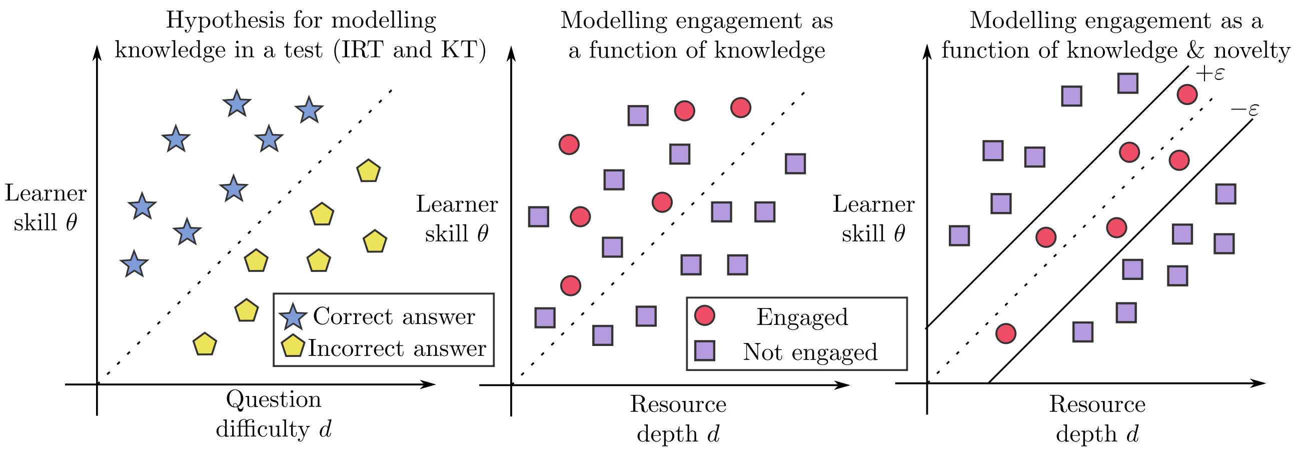

Figure 8: Graphical representation of different assumptions that can be made when modelling learner's knowledge. The methods tested in this paper are set to test these three hypotheses.

References

[1] United Nations. Sustainable development goal 4. https://sustainabledevelopment.un.org/sdg4. Accessed: 2019-06-01.

[2] I. Elaine Allen and Jeff Seaman. Online nation: Five years of growth in online learning. Technical report, 2007.

[3] UNESCO. Open educational resources (oer). https://en.unesco.org/themes/building-knowledge-societies/oer. Accessed: 2019-04-01.

[4] William and Flora Hewlett Foundation. Open educational resources. https://hewlett.org/strategy/open-educational-resources/. Accessed: 2019-06-01.

[5] David Wiley and John Levi Hilton III. Defining oer-enabled pedagogy. *The International Review of Research in Open and Distributed Learning*, 19(4), Sep. 2018.

[6] Stefan Slater, Ryan Baker, Jaclyn Ocumpaugh, Paul Inventado, Peter Scupelli, and Neil Heffernan. Semantic features of math problems: Relationships to student learning and engagement. In *Proc. of Int. Conf. on Educational Data Mining*, 2016.

[7] Anthony F. Camilleri, Ulf Daniel Ehlers, and Jan Pawlowski. *State of the art review of quality issues related to open educational resources (OER)*, volume 52 S. - JRC Scientific and Policy Reports of *Publications Office of the European Union 2014*. 2014.

[8] ISO. *ISO/IEC 40180:2017(en): Information technology — Quality for learning, education and training — Fundamentals and reference framework*, 2017 (accessed November 3, 2018).

[9] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality*, 2(3), December 2011.

[10] Morten Warncke-Wang, Dan Cosley, and John Riedl. Tell me more: An actionable quality model for wikipedia. In *Proc. of Int. Symposium on Open Collaboration*, WikiSym '13, 2013.

[11] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proc. of HLT*, 2011.

[12] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*, 2016.

[13] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*. ACL, 2008.

[14] Daniel H. Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. A general multiview framework for assessing the quality of collaboratively created content on web 2.0. *Journal of the Association for Information Science and Technology*.

[15] Sahan Bulathwela, Emine Yilmaz, and John Shawe-Taylor. Towards Automatic, Scalable Quality Assurance in Open Education. In *Workshop on AI and the United Nations SDGs at Int. Joint Conf. on Artificial Intelligence*, 2019.

[16] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. Quality-biased ranking of web documents. In *Proc. of ACM Int. Conf. on Web Search and Data Mining*, 2011.

[17] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proc. of Int. Conf. on World Wide Web*, 2006.

[18] Arnaud Gaudinat, Natalia Grabar, and Celia Boyer. Machine learning approach for automatic quality criteria detection of health web pages. In *Proc. of MEDINFO*, 2007.

[19] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proc. of Int. Conf. on Data Mining*, volume 8, pages 263–272. Citeseer, 2008.

[20] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine.

[21] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proc. of ACM Conf. on Recommender Systems*, 2016.

[22] Andy Lane. Open information, open content, open source. pages 158–168.

[23] Susan Bull and Judy Kay. Smili : a framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, 26(1):293–331, 2016.

[24] Qian Zhao, F Maxwell Harper, Gediminas Adomavicius, and Joseph A Konstan. Explicit or implicit feedback? engagement or satisfaction?: a field experiment on machine-learning-based recommender systems. In *Proc. of the 33rd Annual ACM Symposium on Applied Computing*, pages 1331–1340. ACM, 2018.

[25] Andrew S. Lan, Christoph Studer, and Richard G. Baraniuk. Time-varying learning and content analytics via sparse factor analysis. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 452–461. Association for Computing Machinery, 2014.

[26] G.E. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*, volume 1. 1960.

[27] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[28] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 1994.

[29] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*, pages 171–180. Springer, 2013.

[30] Radek Pelánek. Applications of the elo rating system in adaptive educational systems. *Computers Education*, 98:169 – 179, 2016.

[31] Robert V Lindsey, Mohammad Khajah, and Michael C Mozer. Automatic discovery of cognitive skills to improve the prediction of student learning. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1386–1394. Curran Associates, Inc., 2014.

[32] Devendra Singh Chaplot, Christopher MacLellan, Ruslan Salakhutdinov, and Kenneth Koedinger. Learning cognitive models using neural networks. In *International Conference on Artificial Intelligence in Education*, pages 43–56. Springer, 2018.

[33] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[34] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 505–513. Curran Associates, Inc., 2015.

[35] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. In *Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD)*, 2017.

[36] Arpad Elo. *The Rating of Chessplayers, Past and Present*, volume 3. Arco publishers, London, 7 1978.

[37] Radek Pelánek, Jan Papoušek, Jiří Řihák, Vít Stanislav, and Juraj Nižnan. Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, 27(1):89–118, Mar 2017.

[38] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press, January 2007.

[39] J. Derek Lomas, Kenneth Koedinger, Nirmal Patel, Sharan Shodhan, Nikhil Poonwala, and Jodi L. Forlizzi. Is difficulty overrated?: The effects of choice, novelty and suspense on intrinsic motivation in educational games. In *Proc. of CHI*, 2017.

[40] Sami Abuhamdeh and Mihaly Csikszentmihalyi. The importance of challenge for the enjoyment of intrinsically motivated, goal-directed activities. *Personality and Social Psychology Bulletin*, 38(3):317–330, 2012.

[41] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, 1(1), 2014.

[42] Robert M Carini, George D Kuh, and Stephen P Klein. Student engagement and student learning: Testing the linkages. *Research in higher education*, 47(1), 2006.

[43] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Learning latent engagement patterns of students in online courses. In *Proc. of AAAI Conference on Artificial Intelligence*, 2014.

[44] Andrew S Lan, Christopher G Brinton, Tsung-Yen Yang, and Mung Chiang. Behavior-based latent variable model for learner engagement. In *Proc. of Int. Conf. on Educational Data Mining*, 2017.

[45] Weiyu Chen, Carlee Joe-Wong, Christopher G Brinton, Liang Zheng, and Da Cao. Principles for assessing adaptive online courses. In *Proc. of Int. Conf. on Educational Data Mining*, 2018.

[46] Jim Henry and Jeff Meadows. A convergent participation model for evaluation of learning objects. *Canadian Journal of Learning and Technology / La revue canadienne de l'apprentissage et de la technologie*, 28(3), 2002.

[47] C. G. Brinton and M. Chiang. Mooc performance prediction via clickstream data and social learning networks. In *Proc. of INFOCOM*, 2015.

[48] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proc. of Int. Conf. on Knowledge Discovery and Data Mining*, 2002.

[49] Eric Meyerson. Youtube now: Why we focus on watch time, 2012.

[50] Albert Corbett. Cognitive computer tutors: Solving the two-sigma problem. In Mathias Bauer, Piotr J. Gmytrasiewicz, and Julita Vassileva, editors, *User Modeling 2001*, pages 137–147. Springer Berlin Heidelberg, 2001.

[51] Christopher Bishop, John Winn, and Tom Diethe. *Model-Based Machine Learning*. Early access version (`http://www.mbmlbook.com/`), 05 2015. Accessed: 2019-05-23.

[52] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of Int. Conf. on World Wide Web*, 1998.

[53] Philip J. Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proc. of the First ACM Conf. on Learning @ Scale*, 2014.

[54] James R. Williams. Guidelines for the use of multimedia in instruction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 42(20):1447–1451, 1998.

[55] Todd B Kashdan and Paul J Silvia. Curiosity and interest: The benefits of thriving on novelty and challenge. *Oxford handbook of positive psychology*, 2:367–374, 2009.