Common vs. Expert knowledge: making the Semantic Web an educational model

Muriel Foulonneau
Tudor Research Centre,
29, av. John F. Kennedy
L-1855 Luxembourg, Luxembourg
muriel.foulonneau@tudor.lu

Valentin Grouès
Tudor Research Centre,
29, av. John F. Kennedy
L-1855 Luxembourg, Luxembourg
valentin.groues@tudor.lu

ABSTRACT
Model based learning uses models in order to generate educational resources and adapt learning paths to learners and their context. Many domain models are published on the Web through linked data, thus providing a collective knowledge base that can be reused in the educational domain. However, for these models to be usable in an educational context, it should be possible to predict the learning context in which they can be used. A typical indicator of the usability of learning objects is their difficulty. Predicting the difficulty of an assessment item depends both on the construct, i.e., what is assessed, and on the form of the item. In this paper, we present several experiments that can support the prediction of the difficulty of an assessment item generated from a linked data source and the difficulty of the underlying item construct. We analyze the results of a test carried out with choice items (such as multiple choice questions), together with a Web mining approach, in order to provide indicators that the factual knowledge is common knowledge or expert knowledge in a particular population. Our objective is to annotate the semantic models and increase their reusability in an educational context.

Categories and Subject Descriptors
H.1.2 [**Information Systems**]: Models and principles – User/Machine systems *human information processing.*

General Terms
Algorithms, Measurement, Experimentation

Keywords
Linked Data, Semantic Web, educational model, knowledge level, assessment item generation, paradata.

Linked Learning 2012, 17 April 2012, Lyon, France.

1. INTRODUCTION
A key characteristic of educational resources is their inclusion in a learning context, in particular through semantic properties such as Audience (IEEE Learning Object Metadata property) and psychometric indicators (Usage Data in the IMS Question & Test Interoperability specification[1]). However, this information is the result either of a calibration of the learning resources with sample users, or of the estimate of teachers or learners.

Model based learning involves the creation of learning resources from models. Models, such as ontologies represent the knowledge and skills (e.g., domain models). Model based learning is used for intelligent tutors and item generation for instance. Models are usually created manually by domain experts (e.g., DynaLearn[2], APOSDLE[3]). The educational context and the scope of the model are implicit. Models represent a problem or a domain which must be learned by certain types of learners at a certain stage. They are created with the relevant level of detail for a particular population.

However, when reusing existing domain models for educational purpose (e.g., Foulonneau, 2011), it is necessary to define the scope of that model and the learning context for which it is well adapted. Many factors have to be taken into consideration for selecting the parts of the model that are useful in particular contexts. They include the level of expertise that the learners are expected to have and gain, i.e., the difficulty and the intended audience for the learning outcome or knowledge construct.

In this context, we define the item construct as the knowledge or skill that is measured by an assessment item. A semantic assertion can be considered a fact that needs to be learned and which may be related to a learning path and a level of expertise in a particular domain. The semantic assertion can then become a knowledge construct that can be learned and will be represented in an assessment item as an item construct.

So far mechanisms are in place to analyze the audience and difficulty of an educational resource (that conveys knowledge), not of the knowledge itself. Our work aims therefore to predict the difficulty and learning context of the constructs that can be

[1] http://www.imsglobal.org/question/

[2] http://hcs.science.uva.nl/projects/DynaLearn/

[3] http://www.aposdle.tugraz.at/

extracted from the model. We used three approaches. The calibration of items generated from the model was carried out with sample users taking an online test. We identified factors of difficulty that are specific to the item through the semantic analysis of the items. . Finally, we investigated whether it was possible to predict the likelihood that a fact is widely known in a given population (common knowledge) through Web mining. We compared the results obtained in all three experiments. In all cases, the experiments only provide indicators for a specific data collection method, since there is no mechanism that would allow inferring directly that a fact is only known by experts (level of expertise that corresponds to the knowledge construct rather than difficulty of the item).

Our hypothesis is that part of the Linked Data Cloud can be used as a learning resource if we can assign relevant paradata to RDF assertions because they can help predict their usability in assessment items and learning resources. It is necessary for instance to identify facts that belong to expert knowledge vs. facts that belong to common knowledge. In this paper, we are suggesting strategies for predicting the difficulty of an assessment item and the type of inference that can be made on the underlying assertions.

2. RELATED WORK

The generation of assessment items from semantic models was investigated in several projects, including DynaLearn (Linnebank et al., 2010). Brown et al. (2005) and Lin et al. (2007) have implemented mechanisms to generate items from WordNet (although not in its RDF form). Papasalouros et al. (2010) generate multiple choice questions from OWL ontologies. Sung et al. (2007) generate semantic models before creating items based on those models. Foulonneau (2011) proposed to reuse knowledge published in the Linked Data Cloud to support the generation of assessment items.

However, none of these works allows defining which items are difficult or in which context they should be given. The difficulty of the item may come from a combination of the complexity of the knowledge construct or domain as learning outcome (e.g., relativity theory) and the forms of the items (e.g., the chance that a learner will know the correct answer in an open question is different from its chance to know the correct answer in a choice item). Assessment related research has essentially focused on measuring the difficulty and learning context at item level, rather than at the level of constructs or learning outcomes, since they highly depend on the learners' path.

The Item Response Theory allows calculating the difficulty of an item as well as its discrimination and pseudo guessing (Reise et al., 2005). IRT is widely used in high stake assessment to support the definition of tests for which the quality of items is critical. It is also used for adaptive testing (Van der Linden et al., 2000).

When items are created individually, it is possible to calibrate them and assign them usage data (e.g., difficulty, discrimination, pseudo guessing). Tests can then be created based on this usage data. However, when generating items from templates, it is possible to predict usage data from the calibration of the item template when the construct is not modified for each item (strong theory, Lai et al., 2009). It is also possible to use many indications on the cognitive aspects of the item in order to predict the difficulty of the item, although there is no complete framework that would allow taking into consideration all dimensions of an item (Gierl et al., 2011). However, when the construct is modified for each item, then they all need to be calibrated (weak theory).

The reuse of models published on the Semantic Web would therefore benefit from the addition of annotations on the graphs on the learning context in which the assertions can be used. It would be possible to derive an annotation from data collected with assessment items (e.g., item difficulty), provided that the population is well described.

3. USAGE DATA AT ITEM LEVEL

In order to obtain an approximation of the level of expertise that corresponds to the construct, two series of choice items were generated with the same construct, the same stem and the same correct answer option. They however use different distractors (i.e., incorrect answer options). If a similar result is found in both types of tests for a given item, then our hypothesis is that the item difficulty might be a valuable indicator of the level of expertise that corresponds to the item construct.

3.1 The generation of assessment items from the Linked Data Cloud

In order to validate the usability of the Semantic Web as a source of knowledge for learning applications, we have developed a system, which generates assessment items from Linked Data (Foulonneau, 2011). It was implemented with a number of target access points, including DBpedia and FreeBase through Sindice[4], which aggregates data from the Semantic Web. It uses links, for instance from DBpedia to the Flickr dataset. The system can support the delivery of Choice items and Match items, in IMS-QTI format (IMS Question & Test Interoperability Specification).

In order to create choice items, the system queries a semantic data source (in SPARQL) and collects data to fill an item template. The data include the stem variable, the correct answer option, as well as the distractors (incorrect answer options) and possibly a formative feedback.

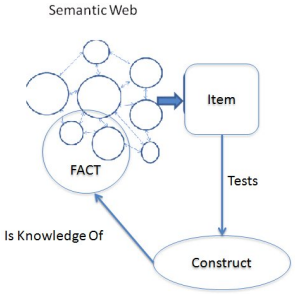

Figure 1 - A semantic fact as a knowledge construct

For each item, instances of the same class as the one of the correct answer option are considered potential distractors.

We extracted 10 facts related to countries and their capital from DBpedia. These facts represent 10 item constructs: does the test taker know that country *x* has a capital named *y*? Two sets of items were generated for each item construct IC, one in test T1, the other in test T2.

[4] http://sindice.com/

Distractors for the items in the test T1 have been randomly selected among a pre-determined number of candidates. For T2, distractors have been selected based on their semantic similarity with the correct answer option.

We added a component (SemSim), which selects the distractors that are semantically closest to the correct answer. The similarity score for two instances is the weighted sum of the taxonomic, relational and attribute similarity metrics defined by (Maedche & Zacharias, 2002). It therefore uses concept properties and data linkage.

The fact extracted from the Linked Data source (e.g., that Einstein was a physicist) is considered as a factual knowledge construct (Figure 1). Items test that a particular test taker has indeed acquired this factual knowledge.

3.2 Item difficulty calculated with IRT

The test was made available online. T1 (the run with random distractors) and T2 (the run with non random distractors) were tested with 46 test takers. 22 test takers answered T1, 24 test takers answered T2. All test takers are French or come from French-speaking countries, although test items were generated in English.

Our objective is to determine whether the test outcomes indicate that certain item constructs created easy items consistently or difficult items consistently, thus indicating a clear trend (easy or not easy).

If the construct is either very easy or very difficult, it is expected that the difference in performance between the items in T1 and T2 will be low. For item constructs IC5 and IC8 for instance, the change of distractors has significantly dropped the rate of correct answers provided by the sample, from 82% to 29% in T1 and from 95% to 33% in T2. However, for the item construct IC7, the change of distractor only dropped the rate of correct answer from 91% to 88%.

Table 1 – Item difficulty in T1 and T2

Item Construct	Estimate T1	Estimate T2
IC1	0.072	-0.288
IC2	-0.298	0.337
IC3	1.961	0.560
IC4	1.755	0.186
IC5	0.084	0.305
IC6	-0.270	-0.349
IC7	-0.737	-1.076
IC8	-1.517	0.751
IC9	-0.284	0.077
IC10	-0.767	-0.502

We used the Item Response Theory to calculate the difficulty of the items in both T1 and T2 according to the Rasch model (Table 1). Higher values represent a higher difficulty of the items. T1 includes 6 items with negative values. Items measuring constructs IC6, IC7 and IC10 have very low values in both T1 and T2. However, items measuring IC3 have a high value in both T1 and T2.

4. FACTORS OF DIFFICULTY OF THE ITEMS

Based on the results obtained from running two tests with distinct items that measure the same constructs, we can set a new hypothesis, i.e., that it could be possible to determine the contribution of the item features (rather than the construct) to the difficulty of the item. By taking into consideration the semantic similarity of distractors, the rate of correct answers was lower than with the random selection of distractors, for all item constructs. It is likely that the semantic similarity of distractors has a direct impact on the difficulty of items. Therefore, it should be possible to predict the contribution of the choice of distractors to the item difficulty from the measure of the semantic similarity of distractors with the correct answer.

4.1 Item dispersity measures

The selection of distractors based on their semantic similarity with the correct answer significantly decreased the rate of correct answers. We set the hypothesis that it can be a relevant predictor of the rate of correct answers.

We therefore created a measure of the *dispersity* (*Di*) of the item, as the average semantic distance between each distractor r and the correct answer a for an item i having n distractors. The dispersity is a float between 0 and 1 (1 being the maximal dispersity).

$$Di(i) = \frac{1}{n}\sum_{i} d(r_i, a)$$

where:

$$d(a, b) = 1 - semsim(a, b)$$

The semantic distance between two elements is defined as one minus the semantic similarity, $semsim$ between those two elements. This semantic similarity is computed by our SemSim component resulting in a float value between 0 and 1, 1 for two identical elements, 0 for two elements with no similarity found.

This measure however does not take into consideration the semantic distance between both distractors. If an item is created with *Jane Austen* as the correct answer and 3 options as *Jane Doo*, *Jane Austen*, and *Julie Austen*, both distractors are close (not semantically but with using a soundex or a string similarity measure). A user could deduce the correct answer from the fact that both distractors are close to the correct answer but not close to each other. An alternative would be to calculate the *global dispersity* (*GDi*) of the item that would include the distance between all options of the items, rather than only between distractors and the correct answer.

$$GDi = \frac{2}{n(n+1)}\left(\sum_{i}^{n}\sum_{j=i+1}^{n} d(\eta_i, \eta_j) + \sum_{i}^{n} d(\eta_i, a)\right)$$

However, if the question is related to a writer (e.g., *who is the author of the novel 'Pride and prejudice'*?), and alternatives include only one writer or only one writer from the XIXth century, then even with a very vague knowledge on the topic, the candidate might guess the correct answer. It is therefore likely that both a high *Di* and a high *GDi* would decrease the difficulty of an item.

4.2 Item dispersity in T1 and T2

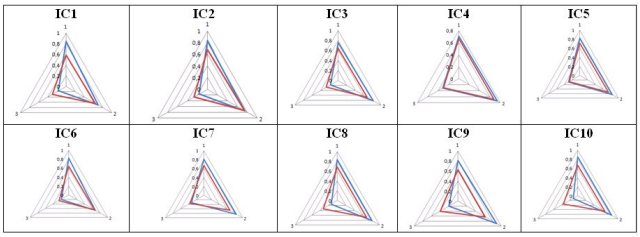

Figure 2 - Representation of the dispersity of the items (T1 in blue, T2 in red)

Figure 2 presents the *dispersity* of items used in our experiment, both when the semantic similarity algorithm was used for the selection of distractors and when it was not used. Axis 3 represents the distance between distractors, whereas axis 1 and 2 represent the distance between a distractor and the correct answer. *Di* only uses Axis 1 and 2, whereas *GDi* uses all three axis. For IC4, IC5 or IC7, the difference is very small between T1 and T2. However, for IC1, IC8, IC9 or IC10, the shapes are significantly different. Without surprise, the distance between distractors and the correct answer (axis 1 and 2) are shorter for T2. However, the distance between distractors themselves tends to be higher for T2. T2 includes more balanced items (triangles closer to equilateral shape), although this parameter was not taken into consideration when selecting distractors for T2.

4.3 Correlation between item difficulty and item dispersity

We calculated the *dispersity* of items in both T1 and T2. Unsurprisingly, the semantic similarity algorithm significantly reduces the dispersity of items (from 0,05 for IC4 to 0,22 for IC9). The Pearson correlation between the percentage of correct answers and the *dispersity* (*Di*) of items is only 0,55 for T1 and -0,43 for T2. Similar results are found for GDi (0,54 and -0,48).

These results would however need to be refined by focusing on cases when the similarity of distractors is expected to have had the most significant impact. However, they open interesting perspectives for the prediction of the contribution of item specific features to the complexity of the item and therefore the identification of the difficulty of constructs derived from knowledge facts published on the Semantic Web.

5. PREDICTING COMMON KNOWLEDGE BY MINING THE WEB

The collection of data from test takers however requires a pseudo-calibration phase, as presented in section 4. This implies a cold start issue similar to the one encountered with recommender systems. As long as a sufficient number of persons has not taken a test item extracted from an RDF assertion, and possibly various forms of test items (e.g., choice items and match items), it is impossible to derive any information on the RDF assertion itself.

We set the hypothesis that for certain types of assertions, it is possible to derive relevant information from the expected exposure of learners to the corresponding concepts. We aim to verify whether we could successfully predict the test results presented in section 3.

In the case of countries and capitals, we verified whether there was a correlation between the Web presence of the concepts and our preliminary conclusions from the user test, on common vs. uncommon knowledge.

We launched searches on the stem variable (country name) and on the correct answer (capital name) on Google. We set the Google advanced search parameters so that only French speaking pages would be returned. A moderate filter was set on the results, the "instant search" parameter was not activated. Google returns a prediction on the number of results, which explains the round numbers provided.

For searches in Google France, we used the French spelling of countries and capitals, although the test was given in English, with two countries having a different spelling in French and in English, and one capital having a different spelling in French and in English.

The Pearson correlation between T1 and T2 and the occurrences of the correct answer options (name of capitals) in Google France is rather high (0,72). This suggests that even a relatively simplistic Web mining approach can provide a satisfactory prediction of the difficulty that can be attached to the RDF assertion. The same experimentation was repeated without specification on the language. The correlations appeared much lower.

In this approach, we assume that the Web presence reflects to some extent the exposure and the familiarity of the population considered with the concepts, this suggests an indirect link between the user knowledge and the correct answer. Indeed, it is not clear whether the exposure to a concept would help predict common knowledge in a particular population more accurately than the exposure to all concepts in the related domain (e.g., geography of Africa) for instance. The prediction of factual knowledge should be considered in light of the other facts already known by candidates and the general awareness of the test takers on a particular topic (culture). Nevertheless, the presence of concepts on the Web can provide an initial insight on the common knowledge acquired by people of a certain culture.

This evaluation should still be consolidated, in particular in order to further characterize the presence of concepts on the Web, and more importantly of combinations of concepts, and taking into consideration the language bias that might have been introduced in this initial evaluation. Nevertheless, these preliminary conclusions suggest that the Web mining approach could be used as one component for the prediction of the difficulty of the learning outcome or the the level of expertise that usually includes this type of knowledge. It should then be possible to use the Semantic Web and the links between datasets to support the construction of learning paths and develop new interactions between users and the Semantic Web.

6. CONCLUSION

In this paper, we showed various strategies to predict the value of an assertion published on the Semantic Web in an educational context. They provided indications, for instance that the item construct IC7 is common knowledge with the population under consideration (French or from French speaking countries), although other cases are more surprising. Although they provided promising results, none of these strategies is in itself sufficient. Further research is needed to refine each of those indicators and assess them as predictors of the difficulty of the generated items.

The annotations of assertions published on the Semantic Web in RDF format are necessary. It is the only way of delimiting semantic subgraphs that can be matched to the learning requirements of a candidate. The target learners and the target learning context should be assessed through different factors of prediction of item difficulty. The item creation mechanisms require developing specific tools, in particular for the assessment of the semantic similarity between concepts, in a meaningful manner for educational purpose. It uses in particular the linkage between datasets, therefore taking advantage of the interconnection between data sources.

This work aims to build mechanisms for making the Semantic Web an educational resource, and therefore to contribute to enhance the Semantic Web with annotations for educational purpose.

Future work will be dedicated to test our approach with teachers, to refine each of the predictors described in this paper, to investigate optimal strategies for combining them, finally to develop a framework for the annotation of the Semantic Web for educational purpose.

7. ACKNOWLEDGMENTS

The present work is partially funded under CLAIRVOYANT core project (FNR C09/IS/12) and supported by the National Research Fund, Luxembourg. We also would like to thank Franck Gismondi and Hélène Mayer for the psychometric analysis.

8. REFERENCES

[1] Brown, J. C., Frishkoff, G. A., & Eskenazi, M. Automatic question generation for vocabulary assessment. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 819–826). (2005)

[2] Foulonneau, M. (2011). Generating Educational Assessment Items from Linked Open Data: the Case of DBpedia. In R. Garcia-Castro et al. (Eds.): ESWC 2011 Workshops, LNCS 7117, pp. 16--27. Springer, Heidelberg

[3] Gierl, M.J., Lai, H. (2011). The Role of Item Models in Automatic Item Generation. Paper Presented at the Symposium “Item Modeling and Item Generation for the Measurement of Quantitative Skills: Recent Advances and Prospects” Annual Meeting of the National Council on Measurement in Education New Orleans, LA April, 2011

[4] Lai, H., Alves, C., & Gierl, M. J. (2009). Using automatic item generation to address item demands for CAT. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

[5] Lin, Y.C., Sung, L.C., and Chen, M.C. An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding. Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007), pages 137-142. (2007)

[6] Linnebank, F., Liem, J., and Bredeweg, B. Question generation and answering. DynaLearn, EC FP7 STREP project 231526, Deliverable D3.3. (2010)

[7] Maedche, A., & Zacharias, V. (2002). Clustering Ontology-Based Metadata in the Semantic Web. *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 348-360). London, UK: Springer-Verlag.

[8] Papasalouros A., Kotis K., & Kanaris K. (2010). Automatic generation of tests from domain and multimedia ontologies. Interactive Learning Environments.

[9] Reise, S.P., Ainsworth, A.T., and Haviland, M.G. (2005). Item Response Theory Fundamentals, Applications, and Promise in Psychological Research. In *Current Directions in Psychological Science April 2005 vol. 14 no. 2 95-101*

[10] Sung, L.-C. Lin, Y.-C., Chen, M. C. The Design of Automatic Quiz Generation for Ubiquitous English E-Learning System. Technology Enhanced Learning Conference (TELearn 2007), pp. 161-168, Jhongli, Taiwan. (2007)

[11] Van Der Linden, W.J., Glas, C.A.W. (2000). Computerized adaptive testing: Theory and practice. Kluwers Academic Publishers, Netherlands.