Apprentissage : corriger et visualiser

Gabriel Illouz[1] Anne-Laure Ligozat[2] Frédéric Vernier[1]

[1] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay
[2] LIMSI, CNRS, ENSIIE, Université Paris-Saclay

Abstract. [De Landsheere, 1980] pose les 3 rôles de l'évaluation : de pronostique, de jaugeage, de diagnostique. Des travaux effectués au LIMSI concernant ces trois aspects sont présentés : l'aide à la correction d'évaluation formatives et sommatives et au suivi d'apprenants, avec une application à la plateforme `Moodle`, des prototypes pour représenter un apprentissage par rapport à une représentation d'un domaine de connaissance.

1 Introduction

Cet article étudie quatre pistes de recherche visant à faciliter l'utilisation d'environnements numériques de travail tout au long d'un parcours d'apprentissage: l'aide à la conception de QCM, l'aide à la correction de réponses courtes, l'utilisation d'ontologies pour structurer l'évaluation, et la visualisation du suivi des apprenants. Nous présentons des systèmes en cours de développement illustrant ces quatre pistes.

[De Landsheere, 1980] souligne l'importance de l'évaluation et des problèmes qu'elle soulève. Il propose trois rôles de l'évaluation : pronostic (savoir si un apprenant a le niveau nécessaire pour réussir un nouvel apprentissage), jaugeage (progrès dans le temps pour un même apprenant, par rapport à un groupe pour une classe), et diagnostic (trouver les causes d'un échec d'apprentissage). Nous présentons dans cet article nos travaux concernant ces trois aspects, qui ont fait l'objet de développements sous la plateforme d'apprentissage en ligne `Moodle` [3]. Notre objectif avec ces travaux est double : améliorer l'expérience des enseignants et des apprenants, en tirant parti des travaux de recherche en traitement automatique des langues et visualisation d'information ; et utiliser la plateforme pour recueillir des données pour améliorer le système d'aide à l'évaluation.

2 Contexte pédagogique applicatif

Nous présentons dans cette section les fonctionnalités et utilisations actuelles de la plateforme `Moodle`.

[3] https://moodle.org/

2.1 Évaluation automatique

`Moodle` permet d'utiliser plusieurs types de questions avec évaluation automatique. Ce type de questions a l'avantage de fournir un retour immédiat pour l'apprenant sur ses erreurs, voire de lui permettre de s'améliorer en recommençant l'évaluation autant de fois que nécessaire. Cela diminue aussi la charge de travail de l'enseignant en temps de correction. Les principaux types de questions de cette catégorie sont les Questionnaires à Choix Multiple, les réponses numériques et les codes en langages formels.

Les Questionnaires à Choix Multiple [4] (**QCM**) permettent de tester que les apprenants ne font pas de confusion et reconnaissent la bonne réponse. Cependant, les QCM sont difficiles à concevoir car ils doivent évaluer effectivement les connaissances des apprenants, ce qui implique par exemple que les réponses incorrectes soient suffisamment proches de la réponse correcte pour tester les possibilités de confusion, mais suffisamment différentes également pour qu'une seule réponse soit correcte.

Les réponses **numériques** permettent de valider que les apprenants savent retrouver une réponse numérique, sans garantie cependant que le raisonnement soit correct.

Les **codes en langages formels** en programmation ou en mathématiques (sous réserve d'utiliser un formalisme connu) permettent grâce à des modules existants, par exemple VPL [Thiébaut, 2015] d'exécuter du code et de l'évaluer grâce à des jeux de tests ou encore des scripts. Cela permet d'évaluer les apprenants lorsqu'ils produisent un code.

Ces trois types d'évaluation sont limités, puisqu'ils ne permettent pas d'évaluer par exemple la qualité d'un raisonnement ou la capacité à poser une définition.

Il est donc nécessaire d'évaluer les apprenants avec d'autres types de questions, notamment des réponses de type composition courte (une ou plusieurs phrases). Les avancées actuelles en traitement automatique des langues nous permettent en outre d'envisager de les utiliser pour aider le correcteur.

2.2 Évaluation manuelle : composition courte

Les questions ouvertes de type composition courte peuvent être des questions de définitions (*"Qu'est-ce qu'une classe abstraite en Java ?"*), d'explication (*"Comment obliger les valeurs d'une colonne à être uniques en SQL ?"*) ou de justification (*"Pourquoi ne peut-on insérer le tuple (1,3,5,3,2) ?"*).

Pour corriger celles-ci, `Moodle` ne propose qu'une correction manuelle. La notation consiste en un champ texte de commentaire libre et une note. Cela pose plusieurs soucis : le temps de correction est important ; il n'est pas possible de garantir la consistance de la correction (une même note pour une même réponse), en particulier s'il y a de nombreux apprenants ; enfin, les commentaires doivent être dupliqués pour toutes les réponses identiques.

[4] Nous ne distinguons pas les différents types : choix multiple/unique, vrai/faux...

2.3 Suivi de la progression des apprenants

Les évaluations ne fournissent qu'une évaluation chiffrée des compétences et connaissances des apprenants, mais pour avoir des retours plus fins, il est intéressant d'étudier la progression des apprenants. `Moodle` propose un suivi de compétences au niveau des tests grâce à un référentiel, mais cela correspond à des retours assez globaux, les tests comprenant des questions sur différentes connaissances. On peut considérer qu'une compétence (par exemple : écrire un rapport) nécessite un certain nombre de connaissances (en orthographe, grammaire, sur le sujet traité).

Néanmoins, cela est peu adapté à nos usages. En effet, un référentiel de compétences ne représente pas forcément les pré-requis et les chemins possibles dans un apprentissage. La validation de compétences par un test ou une activité nous semble non pertinente pour valider individuellement des connaissances. En général, nos évaluations portent sur plusieurs concepts.

En tant qu'enseignants, nous avons souligné les problèmes que posent nos usages : la difficulté de concevoir des évaluations appropriées au niveau des apprenants, le temps de correction des évaluations lorsqu'elles nécessitent une correction manuelle, et le manque de suivi fin des apprenants. Nous avons mené plusieurs expérimentations et travaux de recherche qui cherchent à répondre à ces problèmes, et que nous présentons maintenant.

3 Aide à la conception d'examens

Il est difficile de concevoir des QCM qui évaluent les connaissances des apprenants sans biais dus à des défauts de conception (consigne peu claire, formulation négative, réponses non homogènes...). Des outils d'aide à la conception de QCM ont été développés au Limsi [Pho *et al.*, 2015], qui demandent à être intégrés aux plateformes de formation. Nous avons notamment développé une méthode de validation des distracteurs (réponses incorrectes), qui permet d'évaluer si les distracteurs proposés par l'enseignant sont intéressants, ou d'en générer. Cette méthode part de l'analyse des documents du domaine (typiquement, documents de cours) et est fondée sur la consigne d'homogénéité des réponses entre elles : les réponses sont comparées par différentes mesures d'homogénéité sémantique, qui sont combinées. Ces mesures font intervenir des informations issues des textes et des ressources structurées. L'évaluation a montré que la combinaison des mesures choisies obtient de meilleures performances que l'état de l'art. Cependant, ce système présente certaines limites: la méthode est pour l'instant limitée à certains types de réponses (groupes nominaux ou entités nommées), n'a été évaluée que pour l'anglais, et n'a pas été évaluée par des experts. Il reste donc de nombreuses questions de recherche à traiter, à la fois en traitement automatique des langues, et d'un point de vue expérimental.

4 Aide à la correction d'examens

4.1 Interface d'aide à la correction et au recueil de données

Nous avons défini une interface d'annotation permettant de remplir dynamiquement et réutiliser une grille d'analyse. Ce plugin `Moodle`, développé[5] au LIMSI est présenté en figure 1.

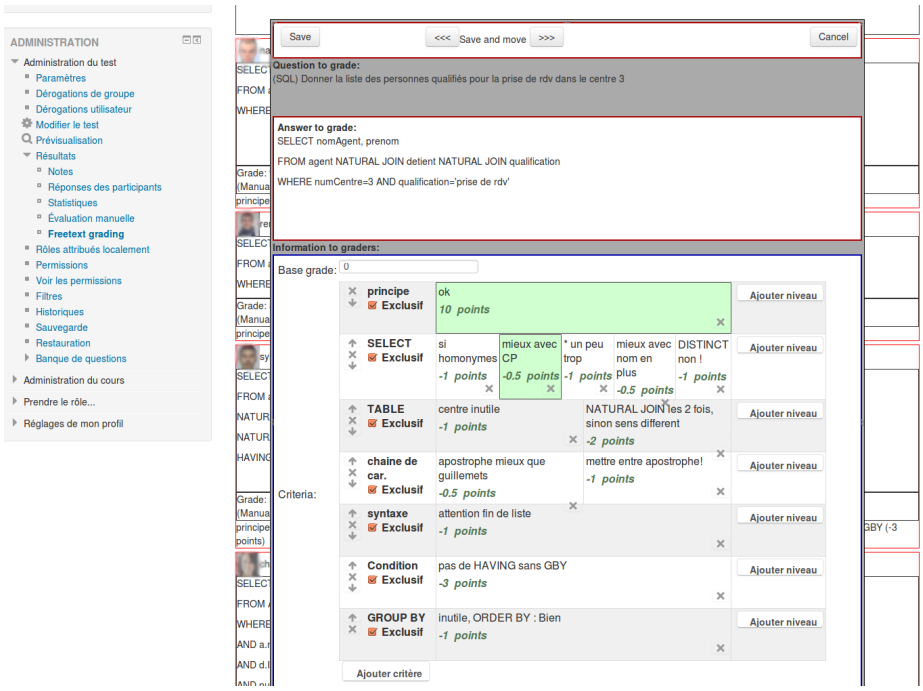

Fig. 1. Interface d'aide à l'annotation

Il permet de n'écrire les commentaires et le nombre de points associés à un cas d'erreur qu'une seule fois. Les premières expériences montrent un gain de temps conséquent (division du temps de correction entre 2 et 4). Cette interface permet également le recueil de données finement (an-)notées, notamment car un retour peut être fait aux étudiants, qui ont la possibilité de contester leurs notes.

Ces données permettent alors d'envisager plusieurs pistes de recherche.

4.2 Correction

Nous souhaitons utiliser les travaux existant en traitement automatique des langues pour faciliter la correction, en regroupant les réponses qui sont proches.

[5] L'objectif est de le distribuer en open source dès que celui-ci sera suffisamment stable.

Dans un premier temps, nous avons développé un module de regroupement de réponses identiques pour `Moodle` : les réponses strictement identiques ne sont ainsi présentées qu'une seule fois à l'enseignant, et les notes et commentaires associés sont propagés à l'ensemble des réponses.

Nous avons également mené des expériences afin d'évaluer les mesures de similarité sémantiques existantes pour le regroupement de réponses proches. Il existe de nombreux travaux sur la correction automatique de réponses courtes ouvertes (voir par exemple [Burrows *et al.*, 2015] pour un état de l'art). Les approches actuelles peuvent être divisées en deux catégories: soit chaque réponse est comparée à une correction donnée par l'enseignant (ou validée en fonction d'indications de l'enseignant, comme par exemple avec des expressions régulières dans le cas d'Open University [6]) ; soit les réponses sont regroupées en fonction de leur proximité entre elles, comme dans [Basu *et al.*, 2013]. Dans tous les cas, une même mesure de similarité textuelle est utilisée pour toutes les questions, et une telle mesure ne pourra jamais atteindre une précision parfaite, ce qui ne satisfait pas le désir légitime des apprenants à être évalué justement. De plus, la même mesure de similarité peut ne pas convenir à la fois à une question de justification et une question factuelle (impliquant une mesure de similarité bien précise) : par exemple l'opposition singlier / pluriel n'est significative que pour certaines questions. Nous souhaitons dans la suite comparer ces approches dans un cadre concret de correction.

4.3 Extraction et structuration de concepts de cours

La création de ressources pédagogiques structurées étant coûteuse en temps, nous avons fait quelques expériences d'extraction et de structuration automatique de concepts à partir de documents de cours [Giannetti, 2013]. Pour cela, nous avons défini un modèle d'ontologie permettant de stocker et structurer des ressources pédagogiques de différents types (définitions, exemples, etc.), et d'adjoindre des ressources terminologiques aux concepts présents dans la taxonomie. Un exemple en est présenté en figure 2. Nous avons ensuite développé des règles destinées à extraire les données pertinentes du corpus et découvrir de nouveaux concepts et mis en place des heuristiques afin de mettre au jour des relations de précédence entre les concepts du domaine.

5 Suivi des apprenants

Les représentations en histogramme sont largement utilisées pour visualiser la distribution des apprenants par rapport aux notes à différents moyennes/contrôles / questions. Néanmoins, cette représentation ne montre pas les aspects temporels et les corrélations, et elle agrège les données sans tenir compte des concepts sous-jacents.

Le découpage en unités minimales de connaissances (notion de concept, ici, pour nous) permet de suivre un élève ou un ensemble d'élèves. Nous proposons

[6] http://www.open.ac.uk/openmarkexamples/

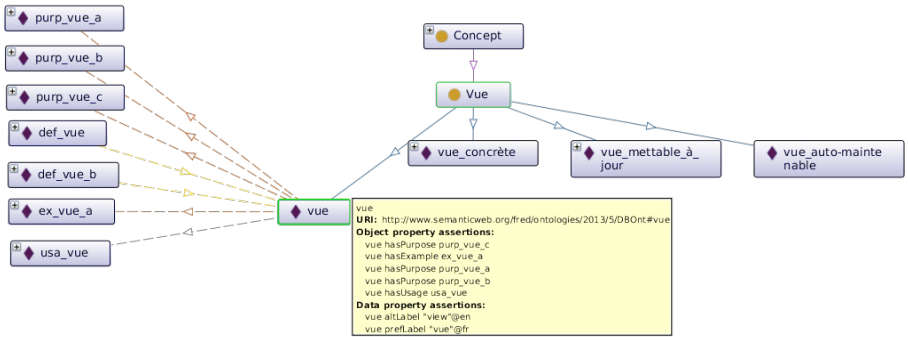

Fig. 2. Exemple d'ontologie pour le concept vue (BD)

une représentation sous forme de tableau avec un code couleur ergonomique (rouge = non acquis, vert = acquis...) pour représenter le niveau de maîtrise d'un concept. La couleur correspond à des intervalles de notes pour un ou plusieurs concepts rencontrés lors d'un ou plusieurs examens.

Les matrices de corrélation permettent d'explorer les corrélations possibles entre de nombreux paramètres (la note à une question, le groupe de TD, la formation d'origine des étudiants...) sans hypothèse a priori.

L'exemple présenté en figure 3, montre les niveaux d'acquisition sur deux concepts testés lors d'une évaluation pour deux étudiants parmi leur groupe. et pour un groupe .

6 Conclusions et Perspectives

Nous avons indiqué dans cet article des développements en cours, les pistes offertes par l'utilisation de travaux de recherche, et les expérimentations que nous envisageons. Notre objectif premier est d'agrandir notre corpus de données d'évaluations annotées. Puis une fois celui-ci obtenu, organiser des défis (challenges) de correction automatique, comme dans le défi ASAP[7] ou encore les données de [Mohler et Mihalcea, 2009], mais en découpant en cas d'erreurs et pas seulement en note.

Nous nous pencherons alors sur l'étude de l'équivalence d'énoncés, en l'envisageant soit comme un problème de reconnaissance de paraphrase [Bouamor, 2012], soit comme un problème d'analogie formelle [Letard, 2017].

La représentation en concepts d'un domaine d'enseignement n'est sans doute pas envisageable de façon totalement automatique. S'aider d'une représentation que l'on puisse manipuler pour arriver à un consensus entre plusieurs enseignants permet d'envisager une ontologie partagée.

[7] https://www.kaggle.com/c/asap-sas

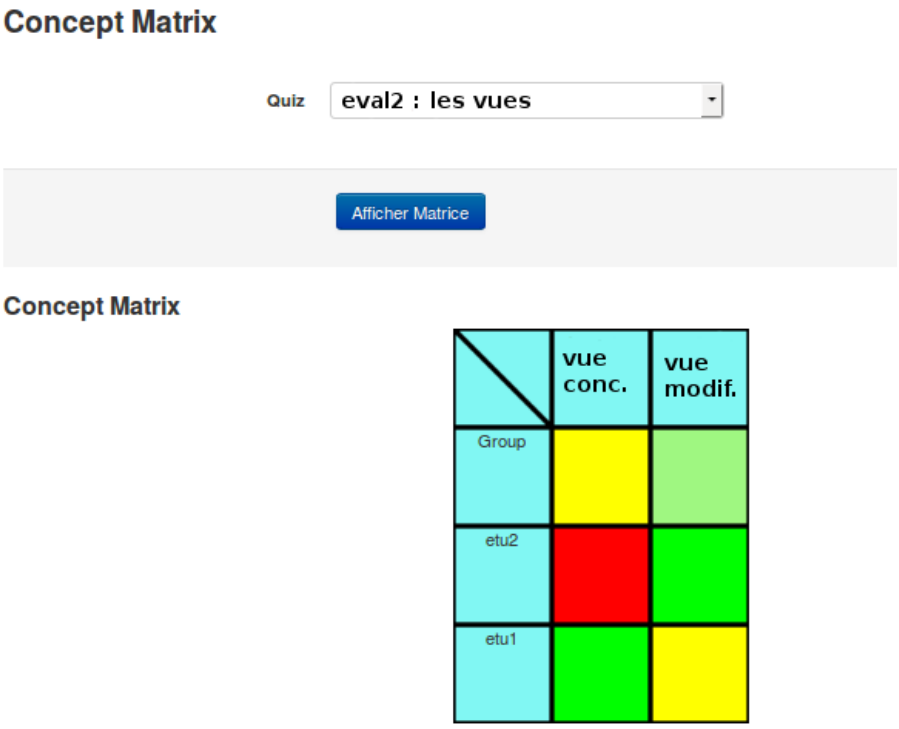

Fig. 3. Exemple de matrice d'apprentissage des concepts

La projection des connaissances et de la progression d'un apprenant sur cet espace de représentations permet d'utiliser cet outil comme une carte de navigation pour aider l'apprenant à atteindre son objectif (un diplôme, ou encore une compétence).

Références

[Basu *et al.*, 2013] Basu, S., Jacobs, C. et Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the ACL*.

[Bouamor, 2012] Bouamor, H. (2012). *Etude de la paraphrase sous-phrastique en traitement automatique des langues*. Thèse de doctorat, Université Paris-Sud – Orsay.

[Burrows *et al.*, 2015] Burrows, S., Gurevych, I. et Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

[De Landsheere, 1980] De Landsheere, G. (1980). *Evaluation continue et examens: précis de docimologie*. Labor.

[Giannetti, 2013] Giannetti, F. (2013). Extraction et structuration de notions de cours pour l'EIAH. Mémoire de D.E.A., Université Paris Ouest – Nanterre – la Défense.

[Letard, 2017] Letard, V. (2017). *Apprentissage incrémental de domaines par interaction dialogique*. Thèse de doctorat, Université Paris-Sud – Orsay.

[Mohler et Mihalcea, 2009] Mohler, M. et Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 567–575. Association for Computational Linguistics.

[Pho *et al.*, 2015] Pho, V.-M., Ligozat, A.-L. et Grau, B. (2015). Distractor quality evaluation in multiple choice questions. *In 17th International Conference on Artificial Intelligence in Education (AIED 2015)*.

[Thiébaut, 2015] Thiébaut, D. (2015). Automatic evaluation of computer programs using moodle's virtual programming lab (vpl) plug-in. *J. Comput. Sci. Coll.*, 30(6): 145–151.