De l'utilisation du dialogue naturel pour masquer les QCM au sein des jeux sérieux

Franck Dernoncourt[1]
(1) LIP6, 4 place Jussieu, 75005 Paris
franck.dernoncourt@lip6.fr

RÉSUMÉ

Une des principales faiblesses des jeux sérieux à l'heure actuelle est qu'ils incorporent très souvent des questionnaires à choix multiple (QCM). Or, aucune étude n'a démontré que les QCM sont capables d'évaluer précisément le niveau de compréhension des apprenants. Au contraire, certaines études ont montré expérimentalement que permettre à l'apprenant d'entrer une phrase libre dans le programme au lieu de simplement cocher une réponse dans un QCM rend possible une évaluation beaucoup plus fine des compétences de l'apprenant. Nous proposons donc de concevoir un agent conversationnel capable de comprendre des énoncés en langage naturel dans un cadre sémantique restreint, cadre correspondant au domaine de compétence testé chez l'apprenant. Cette fonctionnalité est destinée à permettre un dialogue naturel avec l'apprenant, en particulier dans le cadre des jeux sérieux. Une telle interaction en langage naturel a pour but de masquer les QCM sous-jacents. Cet article présente notre approche.

ABSTRACT

Of the Use of Natural Dialogue to Hide MCQs in Serious Games

A major weakness of serious games at the moment is that they often incorporate multiple choice questionnaires (MCQs). However, no study has demonstrated that MCQs can accurately assess the level of understanding of a learner. On the contrary, some studies have experimentally shown that allowing the learner to input a free-text answer in the program instead of just selecting one answer in an MCQ allows a much finer evaluation of the learner's skills. We therefore propose to design a conversational agent that can understand statements in natural language within a narrow semantic context corresponding to the area of competence on which we assess the learner. This feature is intended to allow a natural dialogue with the learner, especially in the context of serious games. Such interaction in natural language aims to hide the underlying MCQs. This paper presents our approach.

MOTS-CLÉS : Agent conversationnel éducatif, intelligence artificielle, jeu sérieux, questionnaire à choix multiple, système d'évaluation de réponses libres.

KEYWORDS : Educational conversational agent, artificial intelligence, serious game, multiple-choice questionnaire, automatic assessment of free-text answer.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 323–336,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Nous définirons dans cette première partie les concepts clés de l'article, nommément le contexte des jeux sérieux ainsi que les agents conversationnels qui constitue la solution que nous explorons pour répondre à la problématique de masquage des QCM.

1.1 Les jeux sérieux

Les jeux sérieux correspondent à une approche de l'apprentissage qui utilise des moyens ludiques. L'apprentissage peut se situer aussi bien dans le cadre d'une formation que dans un contexte de sensibilisation ou de communication (Thomas, 2004). Le marché des jeux sérieux présente une croissance exponentielle : atteignant déjà 1 milliard de dollars en 2004 (Sawyer, 2004), les spécialistes l'estimaient à environ 10 milliards de dollars en 2010.

Dialoguer avec un agent virtuel contribue à maintenir l'attention et la motivation du joueur dans un jeu sérieux. Actuellement, ce dialogue, que ce soit dans les jeux sérieux ou dans les jeux vidéo de type récit (*storytelling*) ainsi que dans la plupart des environnements informatiques pour l'apprentissage humain, est constitué de QCM : le joueur interagit donc avec le jeu avec des QCM, qui font office de dialogue.

Le dialogue est donc très contraint, réduisant ainsi l'apprentissage du joueur qui peut se contenter de cliquer sur une des possibilités sans véritablement réfléchir. Nous pensons que des systèmes de dialogue davantage flexibles peuvent constituer une réponse pertinente à ce problème.

1.2 Les agents conversationnels

Un dialogue est une activité verbale qui fait intervenir au moins deux interlocuteurs servant à accomplir une tâche ou simplement échanger des mots dans une situation de communication donnée. Il constitue une suite coordonnée d'actions (langagières et non-langagières) (Vernant, 1992).

L'idée d'une interaction homme-machine se basant sur le fonctionnement du langage naturel n'est pas nouvelle : elle a vu le jour dans les années 1950 avec le test de Turing. Néanmoins, cette problématique, aux niveaux conceptuel et pratique, demeure toujours d'actualité. Il existe, par exemple, des compétitions annuelles comme le Loebner Prize (Loebner, 2003) ou le Chatterbox Challenge visant à réussir un test de Turing en imitant l'interaction verbale humaine, mais aucun programme n'est parvenu à ce jour à atteindre le niveau d'un humain (Floridi et al., 2009).

Afin de définir des critères d'efficacité des agents conversationnels, nous allons prendre en compte les quatre critères suivants pré-conditionnant l'élaboration d'un système de dialogue intelligent et proposés par (Rastier, 2001) :

1. apprentissage : intégration au moins temporaire d'informations issues des propos de l'utilisateur ;
2. questionnement : demande de précisions de la part du système ;
3. rectification : suggestion de rectifications à la question posée, lorsque

nécessaire ;

4. explicitation : explicitation par le système d'une réponse qu'il a apportée précédemment.

Les agents conversationnels se divisent en deux classes principales :

- les agents conversationnels non orientés tâche destinés à converser avec l'utilisateur sur n'importe quel sujet avec une relation souvent amicale, tel ALICE (Wallace, 2009) ;
- les agents conversationnels orientés tâche, lesquels ont un but qui leur est assigné dans leur conception.

Les agents conversationnels orientés tâche sont eux-mêmes classés usuellement en deux catégories :

- les agents conversationnels orientés service, par exemple fournir un service de conseil sur un site Internet, telle l'assistante virtuelle Sarah de PayPal[1] ;
- les agents conversationnels éducatifs, dont le but est d'aider l'utilisateur à apprendre.

Notre travail se concentre sur les agents conversationnels éducatifs (*tutor bots*).

2 État de l'art

Après avoir posé les définitions de base dans la partie précédente, nous exposerons ici brièvement l'état de l'art sur l'architecture des agents conversationnels ainsi que sur les systèmes d'évaluation des réponses libres plus en détails.

2.1 Architecture d'un agent conversationnel

La figure 1 montre un exemple d'architecture d'un agent conversationnel. L'utilisateur entre une phrase que l'agent conversationnel convertit en un langage abstrait, ici Artificial Intelligence Markup Language (AIML) : cette traduction permet d'analyser le contenu de la phrase et de faire des requêtes via un moteur de recherche dans une base de connaissances. La réponse est générée via un langage abstrait, ici également AIML, qu'il faut traduire en langage naturel avant de la présenter à l'utilisateur.

Néanmoins, cette architecture est rudimentaire et très rigide. Il faut par exemple souvent mettre à jour la base de connaissances pour y inclure des connaissances sur l'utilisateur, notamment dans le cadre d'une activité de tutorat qui nécessite le suivi des acquis de l'utilisateur ainsi que de sa motivation. Un certain nombre d'agents conversationnels éducatifs ont déjà été conçus et implémentés, comme (Zhang et al., 2009), (De Pietro et al., 2005), (Core et al., 2006), (Pilato et al., 2008) ou encore (Fonte et al., 2009).

Diverses architectures ont été élaborées, voici les éléments communs à la plupart d'entre

[1] https://www.paypal-virtualchat.com/

elles :

- une base de connaissances inhérente au domaine, objet de l'application ;
- un gestionnaire de répliques ;
- des structures de stockage des échanges sous forme d'arborescences surtout dans les agents conversationnels éducatifs conçus dans le cadre d'un jeu vidéo.

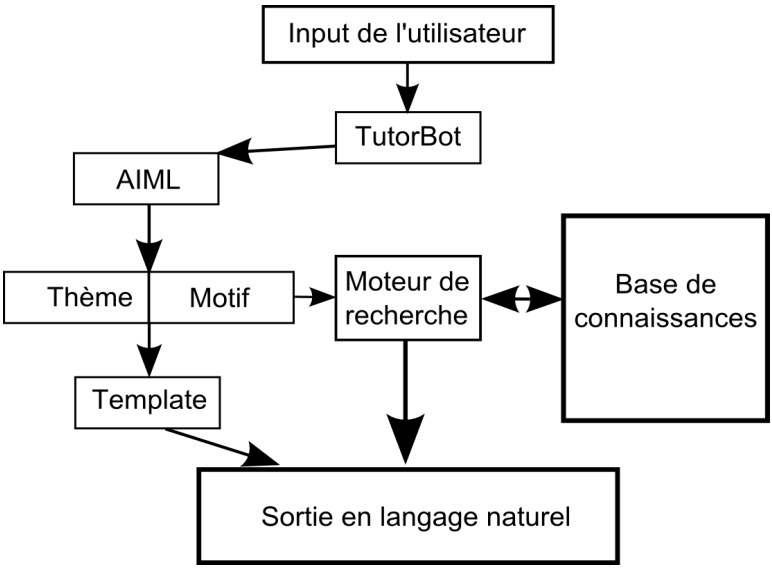

FIGURE 1 – Exemple d'architecture d'un agent conversationnel (TutorBot)
Source : (De Pietro et al., 2005).

Bien que sa simplicité d'utilisation ainsi que la performance relativement bonne des agents conversationnels l'utilisant le rendent attrayant, AIML est un langage très limité qui peut se résumer à un simple filtrage par motif, les motifs des inputs (phrases de l'utilisateur) et des outputs (réponses de l'agent conversationnel) étant définis en grande partie par extension et a priori.

2.2 Systèmes d'évaluation des réponses libres

En parallèle de la recherche sur les agents conversationnels, beaucoup de travaux se sont penchés sur l'évaluation des réponses libres, c'est-à-dire en langage naturel, données par des apprenants. Ces travaux sont motivés par les résultats expérimentaux montrant les limites des QCM, en tant qu'outil d'évaluation de la connaissance des

apprenants (Whittington et Hunt, 1999), ainsi que sa complémentarité avec les réponses libres (Anbar, 1991). Par connaissance, nous entendons ici et dans le reste de l'article non seulement la capacité à retranscrire des informations précédemment apprises, mais également la capacité à opérer des raisonnements de base montrant la compréhension du sujet.

Par exemple, (Anbar, 1991) a montré que les étudiants qui excellent lors des examens oraux auront tendance à avoir des performances médiocres dans les QCM. Inversement, les résultats aux QCM ne permettent pas de prédire les performances de l'apprenant dans le cadre d'un dialogue en langage naturel.

Nonobstant ces limitations bien connues des QCM, ces derniers représentent toujours l'outil le plus utilisé pour les évaluations des apprenants. Ce paradoxe s'explique simplement par le coût beaucoup plus élevé des méthodes alternatives : s'il est trivial de corriger automatiquement les QCM, il n'en va pas de même des autres méthodes, lesquelles nécessitent, étant données les techniques actuelles, des interventions humaines longues, donc coûteuses.

L'évaluation automatique des réponses libres a toutefois également ses détracteurs, qui soulignent que le fait qu'évaluer un essai est une tâche par nature complexe et subjective. Cependant, cette subjectivité ayant pour conséquence une variation de notes non négligeable parmi les correcteurs humains, le système d'évaluation automatique pourra au moins être consistant dans sa subjectivité.

Les premières recherches sur l'évaluation automatique apparurent il y a une cinquantaine d'années. Un des projets remarquables fut le Project Essay Grade, dirigé par Ellis Batten Page à l'université Duke (Page, 1968). Ses travaux se sont basés sur l'utilisation des caractéristiques stylistiques de la réponse de l'apprenant, tels la taille des mots et le nombre de prépositions, pour prédire la note du correcteur humain. Dans ses dernières expériences (Page, 1995), ce système semble prédire la note du correcteur humain plus précisément que ne le fait un second correcteur humain.

À la fin des années 1980, une nouvelle technique a été développée en vue de mieux saisir les concepts sous-jacents à un texte : l'analyse sémantique latente (LSA) (Deerwester et al., 1988 ; Deerwester et al., 1990). Cette technique fut dans un premier temps utilisée dans le cadre de recherche de l'information ; elle ne fut que plus tard appliquée à l'évaluation des réponses libres. La LSA serait aisée à réaliser si un mot ne correspondait qu'à un seul concept, et inversement. Néanmoins, dans les langages naturels, un mot peut avoir différentes significations : un mot peut subséquemment faire référence à différents concepts, faisant ainsi apparaître une ambiguïté à l'échelle du mot. La LSA utilise le contexte dans lequel le mot est utilisé afin de lever l'ambiguïté, autrement dit de comprendre à quel concept le mot fait référence dans le contexte donné. Par exemple, le mot *vol* désigne très certainement le concept de soustraire frauduleusement le bien d'autrui si le mot est utilisé à proximité des mots *butin* et *dérober*. Par contre, si mot *vol* est proche des mots de *ciel* et *oiseau*, *vol* désigne alors probablement un moyen de locomotion aérienne. La figure 2 illustre l'objectif de la LSA.

La LSA ne prend pas en compte l'ordre des mots, ni a fortiori les relations syntaxiques ou logiques. En outre, elle peut s'avérer assez coûteuse d'un point de vue computationnel. Malgré cela, des expériences ont montré que les scores de qualité

globale à un essai donnés par des experts sont moins précis que le score résultant d'une LSA (Landauer, 1998). Ce résultat surprenant est néanmoins à relativiser au vu des limitations de la LSA précédemment mentionnées et dépend évidemment des conditions de l'expérience.

Une approche totalement différente de la LSA a été adoptée par l'Educational Testing Service (ETS). ETS est la plus grande organisation privée à but non lucratif de mesure et d'évaluation éducative au monde. Faisant passer plus de 20 millions d'examens annuellement (TOEFL, GRE, GMAT, etc.), ETS peut ainsi avoir accès à des corpus considérables. Depuis plus d'une vingtaine d'années, son département R&D travaille sur des solutions permettant de noter automatiquement les réponses des candidats. Après avoir essayé d'utiliser la LSA afin de classifier les réponses (Burstein et al., 1996), ETS a décidé de s'en éloigner pour développer la technologie c-rater (Leacock et al., 2003), C pour contenu, qui se focalise sur les réponses de petite taille, allant de quelques à une centaine de mots. C-rater se base sur un pré-traitement de la réponse suivant l'architecture présentée à la figure 3. Ce pré-traitement permet de faire apparaître dans la réponse diverses caractéristiques linguistiques, tels les POS tags, les lemmes de chaque mot ou la présence de négation. Ces caractéristiques linguistiques sont ensuite utilisées pour comparer la réponse du candidat avec une réponse modèle à l'aide d'un algorithme de détection des concepts nommé *Goldmap*. Dans un premier temps, Goldmap était basé sur un ensemble de règles de filtrage par motif déterminées de façon binaire. Bien que cela permettait de comprendre aisément les décisions, la binarité des règles induisait un manque important de flexibilité. Afin de faire face à ce problème, Goldmap adopte à présent une approche probabilistique en se basant sur le principe d'entropie maximale pour la détection des concepts et en intégrant une dizaine de règles ad hoc. Les résultats obtenus semblent prometteurs selon leurs auteurs (Leacock et al., 2003). Cependant, à notre connaissance il n'existe pas à ce jour de test de performance standardisé pour comparer les différents systèmes d'évaluation automatique : il est donc difficile de comparer efficacement les différents systèmes.

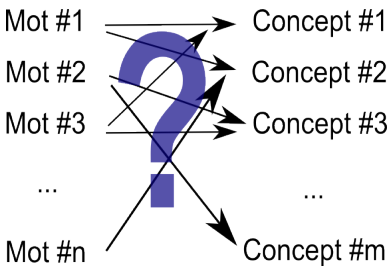

FIGURE 2 – Objectif de la LSA : trouver les concepts auxquels les mots sont associés.

Outre la LSA et c-rater, il est intéressant de noter que beaucoup d'articles soulignent les apports potentiels de la traduction automatique vers l'évaluation de réponses libres. Un des meilleurs exemples est la méthode BLEU (Papineni et al., 2001). Conçue originellement pour évaluer et classer les systèmes de traduction automatique, la

méthode BLEU a été appliquée avec succès à l'évaluation de réponses libres. La méthode repose sur la comparaison entre le texte candidat et un ensemble de textes modèles. Appliquée à la traduction, le texte candidat correspond à la sortie du système de traduction automatique, et les textes modèles correspondent à des traductions réalisées par des experts humains. La note donnée par BLEU au texte candidat se base sur le nombre de N-grammes communs entre le texte candidat et les textes modèles, ce qui s'avère être une mesure efficace malgré sa simplicité, mais est toutefois très sensible au choix d'écriture dans les textes modèles. Lorsque BLEU est appliqué à l'évaluation de réponses libres, le texte candidat correspond alors à la réponse de l'apprenant, et les textes modèles correspondent à des réponses types données par les professeurs. Néanmoins, BLEU présente des limitations importantes, comme par exemple la mauvaise gestion des négations : une phrase niant un fait A aurait par exemple presque le même score qu'une phrase affirmant A.

Au-delà de la méthode BLEU, il est intéressant de remarquer que le domaine de la traduction ainsi que de l'évaluation cherche le même idéal : trouver un formalisme dans lequel les faits pourraient être exprimés indépendamment de langage.

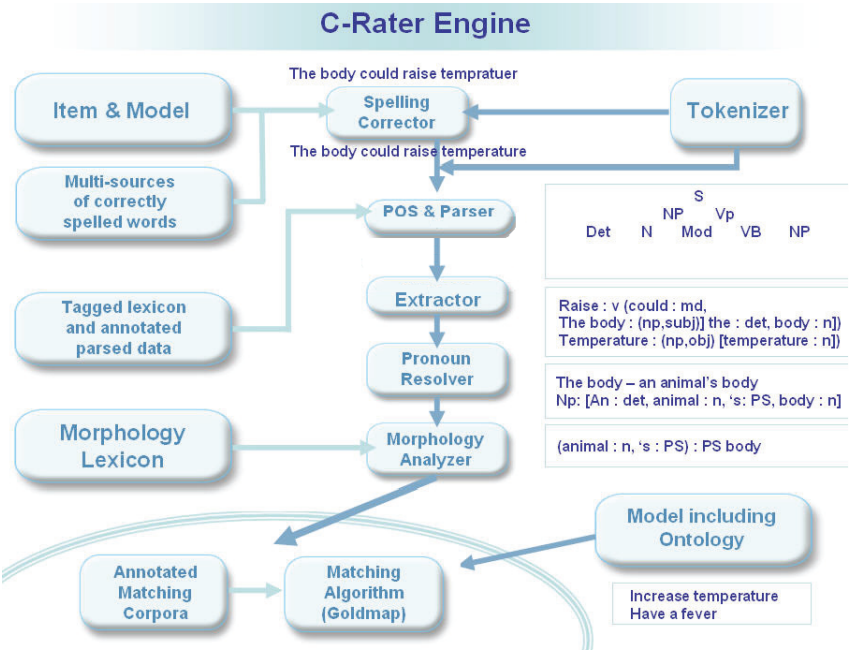

FIGURE 3 – Architecture de c-rater. Source : Sukkarieh et al., 2009.

3 Approche

Nous avons vu dans la partie précédente que beaucoup de travaux se sont penchés sur les systèmes d'évaluation de réponses libres. Dans cette partie, nous mettrons en exergue les particularités de notre approche, en particulier les particularités afférentes à l'évaluation de la réponse libre à l'aune du QCM sous-jacent ainsi qu'à l'environnement de jeu sérieux.

3.1 Particularités des QCM

Nos travaux ont pour but de donner une note à la réponse de l'apprenant. Dans notre approche, nous nous distinguons des systèmes d'évaluation classiques de réponses libres de par deux points majeurs :

- La réponse de l'apprenant n'est pas notée par rapport à des réponses modèles, mais est reliée à un QCM sous-jacent ;
- Une interaction est possible avec l'apprenant, car le système a la forme d'un agent conversationnel.

Ainsi, les recherches se sont penchées sur l'évaluation de réponses libres mais à notre connaissance aucune n'a cherché à évaluer une réponse libre à l'aune d'un QCM sous-jacent. Nous allons donc élaborer des variantes aux techniques habituelles (LSA, BLEU et c-rater) afin de les adapter à l'utilisation de QCM.

L'intérêt d'apparenter la réponse de l'utilisateur à un QCM est multiple. D'une part, de nombreux tests d'évaluation se présentent actuellement sous forme de QCM : nous pourrions ainsi nous baser directement sur les tests existants. D'autre part, la littérature sur la génération automatique de QCM à partir d'ontologie est riche (Papasalouros et al., 2008) : nous pourrons donc à terme avoir un système complet d'évaluation directement à partir des ontologies, voire des supports de cours. Le QCM permet de faire la jonction entre la base de connaissances que constitue le cours et les tests donnés à l'apprenant.

Dans un QCM, l'apprenant choisit une ou plusieurs réponses. Outre les choix corrects, il existe également un certain nombre de choix incorrects. Ces choix incorrects permettent de détecter la présence d'erreur chez l'apprenant de façon active, c'est-à-dire en vérifiant directement si la réponse ne contient pas le choix incorrect. Cette détection active des erreurs est absente de la plupart des systèmes d'évaluation de réponses libres car ces derniers ne reposent que sur la comparaison avec des phrases modèles. Nous pouvons par conséquent identifier ces erreurs alors que les systèmes classiques ont tendance à les ignorer.

Le fait que le système soit sous la forme d'un agent conversationnel nous permet naturellement de faire face plus aisément aux situations où la réponse de l'apprenant ne réussit pas à être directement évaluée par le système : via l'agent conversationnel, une nouvelle question pourra être posée à l'apprenant afin de l'inviter à reformuler ou préciser sa réponse. Cette interaction avec l'agent conversationnel peut être comparée aux tests oraux avec un examinateur humain et permet donc d'éviter les inconvénients émanant des examens écrits classiques qui sont par nature statiques.

3.2 Insertion dans un environnement ludique et sérieux

La simulation d'un dialogue naturel avec le joueur dans un jeu vidéo date d'une trentaine d'années. Le jeu d'aventure *King's Quest I: Quest for the Crown* développé par Sierra On-Line et publié en 1984 figure parmi les pionniers dans le genre. Ce n'est que récemment que la modalité conversationnelle a été utilisée à des fins pédagogiques, notablement dans le jeu Façade (Mateas et al., 2005), que nous allons très brièvement présenter dans le paragraphe suivant.

Dans Façade, le joueur est invité à un dîner où se déroule un conflit marital : l'objectif du joueur est de réconcilier le couple. Pour cela, le joueur entre des phrases de manière écrite, et les deux membres du couples répondent oralement. La figure 4 montre une capture d'écran dans laquelle le joueur demande à la femme Grace si elle se sent énervée vis-à-vis de son mari Trip. En interagissant ainsi avec le couple, le joueur apprend à mieux comprendre les relations de couple.

FIGURE 4 – Capture d'écran de jeu Façade. Le joueur interagit avec le couple.

Néanmoins, jusqu'à présent, ce genre de système de dialogue repose essentiellement sur le repérage de mots clés en fonction desquels le scénario du jeu s'adapte et ne fait pas appel à un QCM sous-jacent. Afin de nous focaliser sur les aspects agent conversationnel et QCM, nous intégrons notre système au sein de la plate-forme Learning Adventure[2] (Carron, 2010).

Learning Adventure est un environnement ouvert en 3D, en ligne et multijoueur où l'apprenant doit réaliser des quêtes en réalisant diverses activités qui le font interagir avec l'environnement et les autres joueurs. L'accent est mis sur le caractère immersif du jeu, à l'instar des MMORPG populaires actuels. L'interaction avec les autres joueurs, autrement dit avec les autres apprenants, est une dimension importante du jeu car elle contribue grandement à la motivation du joueur : le QCM devient pas un jeu solitaire, mais un jeu social, où entrent alors les mécanismes classiques de motivation par les pairs (Dickey, 2007) (Kim et al.,2009).

[2] http://learning-adventure.eu

Outre la motivation résultant de cette collaboration et compétition entre les apprenants, cette dimension multijoueur peut également donner l'occasion pour un tuteur humain d'intervenir dans le jeu. Une telle intervention peut avoir plusieurs objectifs : aider les apprenants dans les tâches réputées difficiles, renforcer les relations élèves-professeur en partageant un moment ludique, etc.

La modalité en ligne du jeu présente quant à elle de nombreux autres intérêts, en particulier s'assurer que le contenu pédagogique est à jour, suivre aisément l'avancement des différents apprenants et faciliter le déploiement de nouveaux contenus.

La figure 5 illustre un QCM qui apparaît dans le cadre du jeu. La figure 6 présente l'éditeur de scénarii, qui permet notamment d'ajouter et modifier aisément des QCM sans avoir aucune compétence informatique particulière. Notre système a pour objectif à terme de rendre le QCM invisible et d'utiliser l'éditeur de scénarii pour permettre à l'enseignant d'inclure les QCM ainsi que les autres éléments du scénario pédagogique.

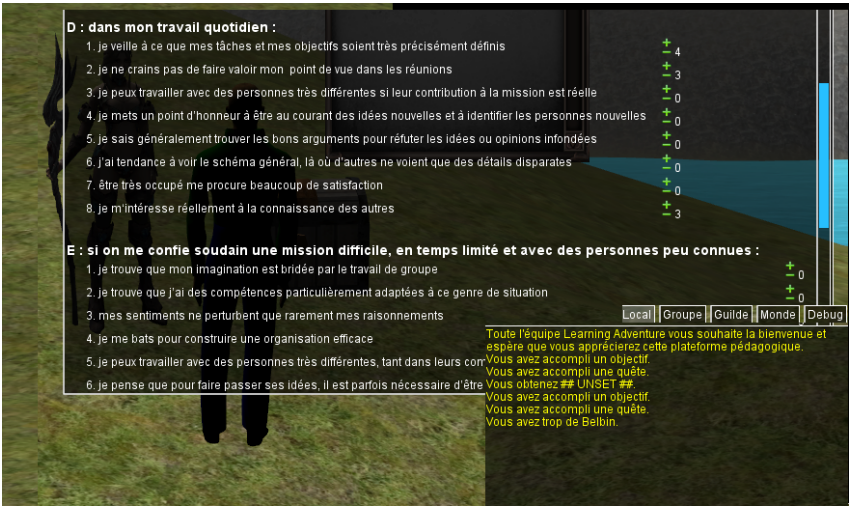

FIGURE 5 – Capture d'écran de la plate-forme Learning Adventure (Carron, 2010)

À l'instar de c-rater d'ETS, nous opérons un pré-traitement de la phrase de l'apprenant, comme le montre la figure 3 : après une étape liminaire de correction de surface (orthographe, ponctuation, etc.), la réponse est analysée lexicalement, puis étiquetée grammaticalement et parsée, avant de passer à l'étape de désambiguïsation référentielle et d'analyse morphologique. Comme ces processus sont supposés être connus chez le lecteur, nous ne les détaillerons pas ici. Sur le plan technique, ces étapes (hormis la première qui utilise le correcteur Hunspell) se basent essentiellement sur la librairie open-source et gratuite NLTK (Bird et al. 2009) qui offre de nombreuses fonctionnalités de traitement des langages.

Une fois la réponse pré-traitée, chaque mot de la réponse est décrit par un ensemble

d'attributs. Nous pouvons alors la comparer avec les différentes réponses du QCM. Les données extraites lors de la phase de pré-traitement contiennent de nombreuses informations essentielles pour la comparaison, telle la présence de négation.

Afin d'accroître la robustesse du système, nous envisageons de regrouper les résultats issus de la comparaison se basant sur la réponse pré-traitée avec les algorithmes statistiques LSA et BLEU. Cette comparaison multiple ainsi que les résultats obtenus seront exposés en détail dans un futur article, le but du présent article étant de présenter globalement les objectifs, le contexte du problème ainsi que la structure générale de notre système.

Les agents conversationnels classiques n'utilisent que très peu les outils usuels du TALN, comme le montre cet exemple suivant issu du site officiel de Cleverbot.com, qui est pourtant un des agents conversationnels les plus réputés :

Human: Hi.

Cleverbot: Are you a boy?

Human: Yes I am a boys

Cleverbot: Are you a girl?

Dans cet exemple, nous voyons que la faute de frappe de l'utilisateur qui a écrit « boys » au lieu de « boy » a manifestement empêché l'agent conversationnel Cleverbot de comprendre la phrase et de proposer une suite cohérente au dialogue.

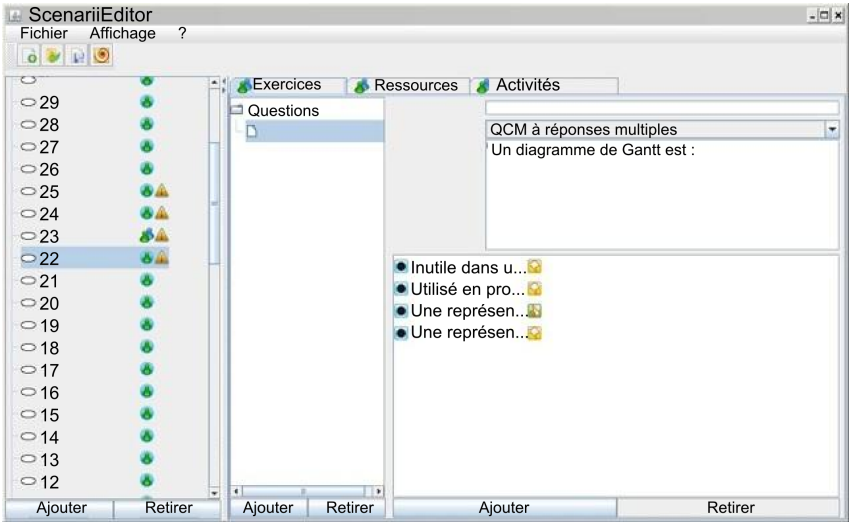

FIGURE 6 – L'éditeur de scénarii pour Learning Adventure

En restreignant le champ sémantique et en précisant son objectif, nous pouvons ainsi intégrer les techniques usuelles du TALN dans notre agent conversationnel afin de rendre transparents les QCM vis-à-vis de l'apprenant.

Enfin, comme le montre (D'Mello et al., 2010), l'apprentissage par agent conversationnel est amélioré lorsque la modalité est orale et non écrite. Par conséquent, nous utilisons Dragon NaturallySpeaking 11, qui est le leader de la reconnaissance vocale et édité par la société Nuance, ainsi que le logiciel AT&T Natural Voices® Text-to-Speech pour transmettre les réponses de l'agent conversationnel sous forme orale. À noter que ces deux logiciels ne sont pas libres.

4 Conclusions et perspectives

Cet article a présenté une approche nouvelle pour évaluer les apprenants en se basant sur des QCM masqués par un agent conversationnel au sein d'un jeu sérieux. La nature interactive du dialogue peut apporter au système d'évaluation une dimension nouvelle, permettant notamment de faire des demandes de clarification (Purver et al., 2003).

Une des difficultés dans la recherche de systèmes d'évaluation de réponses libres est l'absence de benchmarks, absence que certains expliquent par des raisons de propriété intellectuelle (Sukkarieh et Blackmore, 2009). Quelles qu'en soient les raisons, cette lacune est gênante pour la recherche dans le domaine.

Depuis quelques mois, trois initiatives majeures MITx, Coursera et Udacity ont été lancées ; leur objectif est de fournir aux internautes des cours en ligne gratuits, qui ont déjà attiré plus de 100 000 étudiants. Tous trois reposent en grande partie (en plus des tests de programmation dans lesquels le code de l'apprenant est évalué sur un jeu de tests) sur des QCM pour évaluer les apprenants, à défaut de systèmes plus efficaces. Or, ces QCM sont critiqués comme étant une des limites de ce genre de cours en ligne dont l'évaluation est entièrement automatique afin de pouvoir garantir la gratuité vis-à-vis d'un nombre important d'apprenants. La demande de masquage des QCM est donc très importante et continuera de s'accentuer par le nombre croissant de cours en ligne.

Au-delà des contextes d'apprentissage, un tel système pourrait également être utilisé dans d'autres domaines comme l'aide personnalisée, à l'instar de celle fournie par les centres d'appel qui est en général très scriptée, c'est-à-dire suivant des scénarii très peu flexibles, correspondant à un enchaînement de QCM.

5 Remerciements

Je souhaite particulièrement remercier mon directeur de thèse Jean-Marc Labat pour ses précieux conseils, indispensables à la réalisation de ce projet, ainsi que la DGA pour son soutien financier. Je souhaite également remercier Thibault Carron pour ses nombreuses idées ainsi que son aide sur Learning Adventure dont il est un des initiateurs.

6 Références

ALHADEFF, E. (2008). Reconciling Serious Games Market Size Different Estimates. *In Futurlab Business & Games Magazine* - Numéro du 9 avril 2008.

BIRD, S., KLEIN, E. et LOPER, E. (2009). Natural Language Processing with Python. O'Reilly Media.

BURSTEIN, J., KAPLAN, R., WOLFF, S. et LU, C. (1996). Using Lexical Semantic Techniques to Classify Free-Responses. *In Proceedings of SIGLEX 1996 Workshop, Annual Meeting of the Association of Computational Linguistics*, University of California, Santa Cruz.

CARRON T., MARTY JC. et TALBOT S. (2010). Interactive Widgets for Regulation in Learning Games. *The 10th IEEE Conference on Advanced Learning Technologies*, Sousse, Tunisia.

CORE, M., TRAUM, D., LANE, H. C., SWARTOUT, W., GRATCH, J., LENT, M. V. et MARSELLA. S. (2006). Teaching negotiation skills through practice and reflection with virtual humans. *Simulation* 82(11):685–701, 2006.

D'MELLO, S., GRAESSER, A. et KING, B. (2010). Toward Spoken Human-Computer Tutorial Dialogues. *Human-Computer Interaction,* (4):289--323.

DE PIETRO, O., M. DE ROSE et G. FRONTERA. (2005). Automatic Update of AIML Knowledge Base in E-Learning Environment. *In Proceedings of Computers and Advanced Technology in Education.*, Oranjestad, Aruba, August (2005): 29–31.

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T., HARSHMAN, R., LOCHBAUM K. et STREETER, L. (1988). Brevet (US Patent 4,839,853).

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T. et HARSHMAN, R., Indexing by Latent Semantic Analysis. *In Journal of the Society for Information Science*, vol. 41, no 6, 1990, p. 391-407.

DICKEY, M. D. (2007). Game design and learning: A conjectural analysis of how massively multiple online role-playing games (MMORPGs) foster intrinsic motivation. *Educational Technology Research and Development*, 55(3), 253–273.

FLORIDI, L., TADDEO, M. et TURILLI, M. (2009). Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest. *Minds and Machines*. Springer.

LANDAUER, T.K., LAHAM, D., REHDER, B. et SCHREINER, M.E. (1997). How Well can Passage Meaning be Derived Without Using Word Order? A Comparison of Latent Semantic Analysis and Humans, *in Proceedings of the 19th Annual Conference of the Cognitive Science Society*.

LEACOCK, C. et CHODOROW, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and Humanities.* pp. 389-40.

LOEBNER, H. (2003). Home Page of the Loebner Prize - The First Turing Test. http://www.loebner.net/Prizef/loebner-prize.html [consultée le 03/03/2012].

KIM, B., PARK, H. et BAEK, Y. (2009). Not just fun, but serious strategies: Using meta-

cognitive strategies in game based learning. *Computers & Education*, 52(4), 800-810. doi:10.1016/j.compedu.2008.12.004.

MATEAS, M. et STERN, A. (2005). Structuring Content in the Façade Interactive Drama Architecture. *AIIDE*.

PAGE, E.B. (1968). The Use of the Computer in Analyzing Student Essays. *International Review of Education*, 14, 210-224.

PAGE, E.B. (1995). The Computer Moves into Essay Grading: Updating the Ancient Test, *Phi Delta Kappan*, 76(Mar), 561-565.

PAPASALOUROS, A., KOTIS, K. et KANARIS, K. (2008). Automatic generation of multiple-choice questionsfrom domain ontologies. *IADIS e-Learning*, Amsterdam.

PAPINENI, K., ROUKOS, S., WARD T. et ZHU, W. (2001). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on Association for Computational Linguistics*. 311—318.

PEREZ, D., ALFONSECA, E. et RODRIGUEZ, P. (2004). Application of the BLEU method for evaluating free-text answers in an e-learning environment. *In Proceedings of the Language Resources and Evaluation Conference* (LREC).

PILATO, G., ROBERTO P. et RICCARDO R. (2008). A kst-based system for student tutoring. *Applied Artificial Intelligence 22*, no. 4: 283-308.

PURVER, M., GINZBURG, J. et HEALEY, P. (2003). On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*. Springer. 235—255.

RASTIER, F. (2001). Sémantique et recherches cognitives, *PUF* (2e éd).

SUKKARIEH, J. Z. et BLACKMORE, J. (2009). c-rater: Automatic content scoring for short-constructed responses. *Florida Artificial Intelligence Research Society (FLAIRS) Conference*, Sanibel, FL.

SAWYER, B. (2004). Serious Games Market Size. *Serious Games initiative Forum 01-04-2004*.

THOMAS, P., éditeurs (2010). *Actes de RJC EIAH 2010 (Rencontres Jeunes Chercheurs en Environnements Informatiques pour l'Apprentissage Humain)*, Lyon. ATIEF.

VERNANT, D. (1992). Modèle projectif et structure actionnelle du dialogue informatif. *In Du dialogue, Recherches sur la philosophie du langage*, Vrin éd., Paris, n°14, p. 295-314.

WALLACE. , S. (2009). Parsing the Turing Test, The Anatomy of A.L.I.C.E. Springer.

WHITTINGTON, D. et HUNT, H. (1999). Approaches to the computerized assessment of free text responses. *In Danson, M. (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough, UK.

ZHANG, H. L., Z. SHEN, X. TAO, C. MIAO et B. Li. (2009). Emotional agent in serious game (DINO). *In Proceedings of The 8th International Conference on Autonomous Agents and Multi-agent* Systems-Volume 2, 1385–1386.