La Latent Semantic Analysis et Le Macro-Professeur : Exemple d'intégration de l'intelligence artificielle à un système de tutorat intelligent

Julien Mercier
Monique Brodeur

This article describes the technique of Latent Semantic Analysis (LSA), an artificial intelligence algorithm that can be integrated into intelligent tutoring systems (ITS). LSA is a technique that extracts some semantic properties of a text, using mathematical and statistical operations. The technique uses huge amounts of digitized text, integrated in a semantic space that acts as a knowledge base. LSA is used to compare words or texts with respect to their semantic content. The algorithm is first described. Then, the nature of a semantic space and the method to build one is made explicit, starting from an example (a semantic space on the topic of psychology, in French). Finally, an example of a recent ITS using LSA is described; it consists of a system to support the teaching of summarizing skills, named Le Macro-Professeur, which is available on the Web.

Cet article décrit la Latent Semantic Analysis (LSA), un algorithme d'intelligence artificielle pouvant être intégré à des systèmes informatiques de tutorat intelligent. La LSA est une technique qui extrait certaines propriétés sémantiques d'un texte par des opérations mathématiques et statistiques. Elle évolue à partir dune grande quantité de texte numérisé et intégré dans un espace sémantique qui fait figure de base de connaissances. La LSA sert à comparer des mots ou des textes entre eux quant à leur contenu sémantique. L'algorithme est d'abord décrit. Ensuite, la nature d'un espace sémantique et la méthode menant à sa création sont exposées, en menant comme exemple un espace sémantique

Canadian Journal of Educational Communication Vol. 27 No. 3, pages 175-188.

de langue française portant sur la psychologie. Enfin, un exemple de système de tutorat intelligent récent utilisant la LSA est apporté ; il s'agit d'un système de support à l'enseignement des stratégies de rédaction d'un résumé, Le Macro-Professeur, qui est disponible sur le Web .

La Latent Semantic Analysis (LSA)

Dans une perspective large, le domaine de l'intelligence artificielle vise notamment à recréer artificiellement, au moyen de l'informatique, les processus inhérents à l'intelligence humaine. Par conséquent, l'intelligence artificielle peut être utilisée pour développer des systèmes de tutorat hautement spécifiques destinés à favoriser l'apprentissage. Par exemple, des chercheurs de l'université McGill ont développé un système de tutorat pour I'apprentissage des statistiques au troisième cycle universitaire (Frederiksen & Donin, 1999). D'autres projets d'envergure ont trait à l'entraînement de mécanicien en avionique ou à l'entraînement au diagnostic médical (Lillehaug & Lajoie, 1998). La section suivante présente quelques généralités et l'aspect mécanique de la *Latent Semantic Analysis*.

Généralités

La *Latent Semantic Analysis* (Landauer & Dumais, 1996, 1997) est un nouveau type d'intelligence artificielle qui tire l'information d'un texte par des opérations mathématiques et statistiques complexes (Landauer, Foltz, & Laham, 1998). Les connaissances ou informations que ce système tire des textes qui lui sont soumis sont représentées dans un espace sémantique ou *semantic space.* Selon plusieurs auteurs (Burgess, Livesay, & Lund, 1998; Landauer & Dumais, 1997 ; Landauer, Laham, Rehder, & Schreiner, 1997), cet espace sémantique reproduit artificiellement, à maints égards, le modèle connexionniste de représentation des connaissances élaboré par Kintsch (1998) : le modèle de construction-intégration. De plus, l'utilisation de la LSA est tout indiquée pour les recherches considérant le *textbase,* c'est-à-dire les propositions dont un texte est composé (Foltz, 1996). Selon Perfetti (1998), la LSA fait partie de ces outils qui représentent le *nec plus ultra* pour des applications au niveau de la recherche quantitative portant sur les productions écrites.

Aspect mécanique

Les deux principaux aspects de cette technique de LSA seront explicités ici. D'abord, la mécanique mathématique et statistique sous-jacente à la LSA sera développée. Ensuite, l'espace sémantique, qui est le résultat de cette

(1) L'adresse intemet est la suivante : http://lsa.colorado.edu/Macro-Professeur/

mécanique, sera décrit puisque c'est de ses qualités que dépend l'efficacité de la LSA.

Mécanique mathematlque utilisée par la LSA

La mécanique centrale à la LSA est basée d'une part sur l'analyse statistique descriptive des mots contenus dans les textes et, d'autre part, sur l'analyse factorielle de ces statistiques. Elle est basée sur le postulat qu'il existe une structure « latente » sous-jacente à l'utilisation des mots à travers l'ensemble des écrits. Le système considère l'apparition de chaque mot dans son contexte, généralement au niveau de la phrase ou du paragraphe dont il est issu. Limitons-nous pour l'instant à un mot, point de départ des deux étapes principales opérées par la LSA.

Première étape: analyse statistique descriptive. Le système considère et compile toutes les phrases ou plus généralement tous les paragraphes au sein desquels ce mot apparaît. On obtient donc une liste de tous les contextes dont ce mot fait partie. À ce niveau, on peut concevoir la signification du mot comme la moyenne de tous les contextes où il a été retrouvé. À un niveau plus mathématique, on représente le résultat de cette première opération dans une matrice, dans laquelle les rangées représentent tous les mots des textes et les colonnes représentent tous les contextes (lignes ou paragraphes) dans lesquels apparaissent les mots. Dans cette matrice, chaque case contient la fréquence d'occurrence d'un mot dans chaque contexte.

Deuxième étape : analyse factorielle. Par la suite, une deuxième opération intervient, qui témoigne de l'importance de chaque mot dans chaque contexte. Comme on considère chaque contexte où un mot apparaît et que ces différents contextes forment des contraintes mutuelles, cette opération imite le principe de satisfaction de contraintes du modèle de construction-intégration (Landauer, Laham, Rehder, & Schreiner, 1997). C'est une forme d'analyse factorielle *(singular value decomposition* ou *SVD)* qui mène à la réduction des dimensions à laquelle il sera fait allusion plus loin. Cette analyse factorielle conduit à l'élaboration d'une matrice qui fait état de l'estimation très juste de la fréquence à laquelle apparaîtrait un mot dans chaque contexte dans le cas où l'on considérerait une infinité d'occurrence de ce contexte. Ici, la signification d'un mot est basée sur l'ensemble des contextes où le mot apparaît, sur l'ensemble des contextes où ce mot pourrait apparaître comme synonyme, et sur l'importance de ce mot dans les contextes considérés. Cette réduction des dimensions augmente considérablement la puissance de la technique. Par exemple, elle permet de mettre les synonymes en évidence. En somme, la nature de cette signification sémantique n'est pas logique. Elle est

en fait basée sur des relations de similarité contextuelle entre les mots. C'est donc dire que cette signification est construite à partir des autres mots et contextes de la matrice, à l'image du processus associatif décrit dans le modèle de construction-intégration. C'est pourquoi il est important de réduire les dimensions de la matrice afin de conserver les facteurs qui caractérisent de façon prépondérante l'ensemble des mots utilisés. Ces facteurs sont en quelque sorte des continuums qui représentent chacun une dimension de la matrice. En termes plus mathématiques, le sens d'un mot est déterminé par sa position sur chacun des continuums retenus. Par extension, la signification d'un paragraphe ou d'un passage entier est établie selon les mêmes mécanismes. Elle est constituée des moyennes des vecteurs sémantiques des mots que le paragraphe ou le passage contient, indépendamment de l'ordre des mots (Landauer, Laham, Rehder, & Schreiner, 1997). Les mots et les contextes ainsi que les facteurs qui les caractérisent forment ce que l'on appelle un espace sémantique, qui symbolise artificiellement le savoir qu'une personne retirerait de la lecture des textes traités par la machine, suggérant l'analogie avec le *situation mode1* de Kinstch (1998).

Par ces opérations mathématiques, la LSA permet l'analyse des informations sémantiques issues des contextes dans lesquels apparaissent les mots. Cette analyse est le point de départ de plusieurs actions essentiellement comparatives portant sur les productions écrites de sujets humains, tels que des élèves et des étudiants de tous les cycles, du primaire à l'université. D'abord, elle permet d'évaluer le résumé d'un texte par comparaison avec le texte original et d'obtenir de la rétroaction sur les particularités de ce résumé. C'est d'ailleurs ce principe qui est mis à profit dans le logiciel présenté en exemple. Elle permet de plus d'évaluer le degré de convergence de plusieurs travaux sur un même sujet. Elle permet aussi de cerner à partir d'un ensemble imposant de connaissances les éléments-clés à retenir sur un sujet particulier. Ces comparaisons s'opèrent sur les vecteurs associés à chaque texte ou partie de texte. Il est à noter que ces opérations ne peuvent être réalisées que sur des textes faisant appel au même espace sémantique. Les caractéristiques d'un espace sémantique, les considérations nécessaires à son élaboration ainsi que ses principales fonctions seront exposées maintenant.

L'espace sémantique

Un espace sémantique est caractérisé principalement par l'ensemble des mots qu'il contient. L'ensemble du vocabulaire contenu dans les textes auxquels le système a eu accès lors de la construction de l'espace sémantique par les mécanismes décrits précédemment détermine la teneur des liens, des dimensions et des informations que l'on pourra retirer des analyses. Ainsi, le système construit des « connaissances » liées au vocabulaire et aux contextes

émanant des textes qu'on lui soumet. On peut donc choisir de lui faire absorber le contenu d'un domaine particulier comme la biologie (McNamara, Kintsch, Songer, & Kintsch, 1996) ou la psychologie (Mercier, 1999, site web) selon le sujet des écrits que l'on intègre dans le système. On peut aussi considérer les écrits plus usuels et quotidiens tels que ceux que l'on retrouve dans les forums de discussion sur l'Internet. Toutefois, le choix des textes se doit d'être réalisé judicieusement selon le matériel utilisé à travers les analyses. Comme la méthode est basée essentiellement sur le vocabulaire et ses relations avec le contexte, il doit y avoir un lien étroit entre le sujet des textes contenus dans l'espace sémantique et le sujet des textes produits par les humains. Les publications récentes au sujet de la LSA (Foltz, 1996 ; Landauer, Foltz, & Laham, (1998) ; Rehder et al., 1998) suggèrent qu'idéalement, on devrait retrouver dans l'espace sémantique l'ensemble des mots utilisés par les humains dans leurs textes.

Comme la LSA a recours aux statistiques, l'élaboration d'un espace sémantique de qualité nécessite une grande quantité de textes. De façon analogue à la taille de l'échantillon garant de sa représentativité de la population, l'augmentation de la quantité de contextes où l'on retrouve un mot favorise l'élaboration d'une signification juste de ce mot. De plus, l'étendue du vocabulaire de l'espace sémantique est tributaire de la quantité de textes traitée. Les espaces sémantiques construits jusqu'à maintenant renferment de 8300 mots à quelques centaines de millions de mots pour un vocabulaire de 30119 à 92000 mots. L'examen des espaces sémantiques réalisés antérieurement suggère d'autres précautions qui seront énoncées ici, bien qu'elles n'aient pas encore fait l'objet d'expérimentations empiriques avec la LSA.

Il semble que les textes doivent favoriser l'extraction de l'information contextuelle. À cet égard, les livres spécialisés dans un domaine semblent être tout indiqués. En effet, les mots qu'ils renferment font partie intégrante d'un discours organisé conformément aux connaissances des auteurs sur un sujet. De plus, les mots utilisés dans l'exposé des idées sont soigneusement choisis par l'auteur soucieux d'être bien compris. Enfin, la ressemblance de la LSA avec le modèle connexionniste suggère que la qualité de l'espace sémantique, et conséquemment de ses applications, dépend de la qualité de ce qui est traité, à l'image de la qualité d'un apprentissage qui est tributaire de la qualité du contenu abordé.

Le fait que certains mots revêtent un sens particulier selon le domaine auquel ils réferent suggère que les textes traités lors de l'élaboration de l'espace sémantique devraient être spécifiques au domaine des textes analysés. Selon cette considération intuitive, les textes devraient être tirés d'un même corpus de connaissances relativement spécifique, tel qu'une science particulière. Il est toutefois possible que des impératifs très pragmatiques viennent

influencer le choix des textes ; le rassemblement d'une telle quantité de matériel requiert un investissement considérable de moyens. Pour limiter les ressources nécessaires, les chercheurs ont avantage à limiter l'importance des textes traités. En effet, les sources disponibles de documents en langage-machine (ASCII) sont l'Internet et les CD-roms encyclopédiques qui demandent une recherche et un traitement permettant de colliger un corpus de textes. L'alternative consiste à numériser des textes sur papier, ce processus nécessitant des ressources considérables. De plus, les capacités limitées du système informatique imposent de réduire les dimensions de l'espace sémantique. Selon ces observations, la portée réduite des espaces sémantiques existants serait due non pas à des précautions quant à l'efficacité de la technique de LSA mais bien à des contraintes externes de nature pragmatique.

Préalablement au travail du premier auteur, il n'y avait pas d'espace sémantique de langue française disponible. Le site Web de la LSA (http:// mis sur pied par des chercheurs de l'Université du Colorado à Boulder est essentiellement anglophone. Il comporte plusieurs applications de la LSA ainsi que des sources d'informations concernant la théorie et les techniques sur lesquelles la LSA est fondée. Mis à part le matériel anglophone, le site ne contenait qu'un espace sémantique trilingue regroupant des articles de magazines européens rédigés en français, en anglais et en allemand. Un autre projet en développement porte sur les débats parlementaires bilingues du parlement canadien. Comme le logiciel présenté dans cet article est réalisé avec des textes académiques de niveau universitaire en langue française, l'élaboration d'un espace sémantique approprié est de rigueur. La qualité de cet espace sémantique est déterminante puisque la qualité de la rétroaction fournie par le logiciel dépendent exclusivement de ses propriétés. En effet, c'est au sein de celui-ci que les comparaisons sémantiques sont opérées.

Exemple d'espace sémantique

Cette section comporte deux parties. D'abord, les principales caractéristiques de l'espace sémantique sont exposées. Ensuite, un exemple d'analyse est proposé, afin d'illustrer la teneur et les limites des résultats obtenus par les différentes analyses avec la LSA.

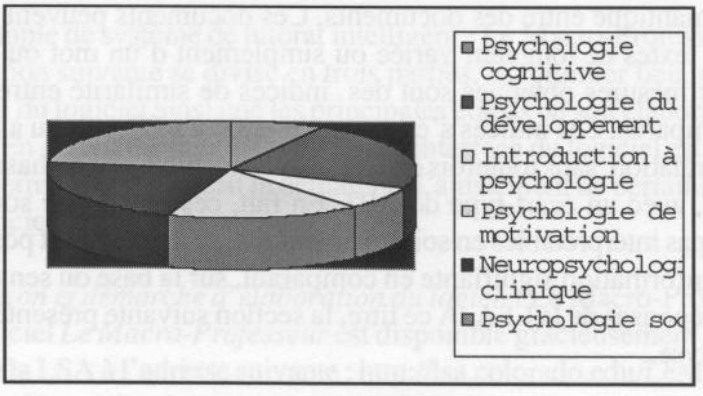
Principales caractéristiques de l'espace sémantique

Cette section comporte des informations sur la nature de l'espace sémantique, sur son élaboration ainsi que sur sa principale fonction. L'espace sémantique présenté dans cette recherche est disponible à la communauté de chercheurs intéressés par des investigations en langue française réalisées avec la LSA, à l'adresse électronique http://lsa.colorado.edu/. Il a pour nom *Psychology_French*.

Nature de l'espace sémantique Psychology_French

L'espace sémantique *Psychology_French* est constitué de 6 livres récents utilisés comme références dans plusieurs cours de première année du programme de baccalauréat spécialisé en psychologie. Les livres font le survol de plusieurs champs disciplinaires de la psychologie : psychologie générale, psychologie cognitive, psychologie sociale, neuropsychologie clinique, psychologie du développement et psychologie de la motivation. La figure suivante offre un aperçu de la proportion occupée par chaque champ de la psychologie au sein de l'espace sémantique.

Figure 1. Proportion occupée par chaque champ de la psychologie au sein de



l'espace sémantique

Élaboration de Psychology_French

Les 6 livres ont été numérisés entièrement, à l'exception des graphiques et des figures. Trois critères principaux ont influencé le choix de numériser des livres plutôt que de l'utilisation de CD-roms ou de textes disponibles en ligne sur Internet. Premièrement, une très grande quantité de livres francophones est disponible. Ensuite, le contenu d'un livre est clairement identifié, notamment par ses descripteurs. Enfin, l'utilisation d'un volume dans le cadre d'un cours universitaire est un indicateur du niveau académique de son contenu. Les résultats de nos recherches sur Internet ou dans les répertoires de CD-Roms n'ont pas répondu à ces trois critères.

Comme le logiciel de reconnaissance de caractères utilisé lors de la numérisation des livres (*Omnipage Limited Edition*) a un taux d'efficacité variant de 30 à 96%, les textes numérisés ont été corrigés à l'aide du correcteur orthographique de *Microsoft Word 95*. Le document *Word* qui a servi de base à la méthode mathématique de la LSA afin de créer l'espace sémantique comporte 2295 pages à interligne simple, soit près de 12 MB de texte. L'espace

sémantique obtenu à partir des livres numérisés comporte un vocabulaire de 41741 mots utilisés dans les champs de la psychologie considérés. À titre de comparaison, *Le Petit Larousse* 1998 contient 59 000 noms communs.

La dimension du contexte a été fixée au paragraphe. Le nombre optimal des dimensions de l'espace sémantique a été fixé à 300. Ce nombre optimal de dimensions permet à la LSA de simuler le jugement humain sur la signification des mots donnant lieu à des résultats de l'ordre de quatre fois la précision d'un espace sémantique sans réduction des dimensions (site web de la LSA).

Fonction d'un espace sémantique

Les analyses menées avec la LSA permettent essentiellement la comparaison sémantique entre des documents. Ces documents peuvent prendre la forme de textes de longueur variée ou simplement d'un mot ou groupe de mots. Les mesures obtenues sont des indices de similarité entre les documents comparés. Ces indices s'étendent de -1.OO à 1.OO, un peu à la manière d'une corrélation, sans toutefois en être une. Deux mots pris au hasard obtiennent 0.02, avec un écart-type de 0.03. En fait, ces indices ne sont, pour le moment, pas interprétables en soi de par leur nature même. Il est possible d'en tirer une information signifiante en comparant, sur la base du sens commun, diverses réponses de la LSA. À ce titre, la section suivante présente un exemple d'analyse.

Exemple d'analyse

Le tableau suivant présente le résultat d'une analyse permettant de comparer plusieurs mots entre eux. Il s'agit de l'application *Matrix Comparison,* que l'on retrouve sur le site Web de la LSA, utilisée avec l'espace sémantique *Psychology_French.*

Tableau 1. Résultat d'une comparaison au sein de l'espace sémantique

	psychanalyse	psycho. développement	béhaviorisme	psycho. sociale
Freud	0.62	0.12	0.04	0.05
Piaget	0.03	0.47	0.07	0.03
Pavlov	0.08	-0.01	0.19	0.07
Vallerand	-0.02	-0.02	0.01	0.02

Pour cette analyse, quatre grands noms de la psychologie ont été entrés, accompagnés des domaines de la psychologie auxquels ils ont

significativement contribué. Freud obtient la meilleure association avec psychanalyse (0.62). Piaget obtient la meilleure cote auprès de la psychologie du développement (0.47), tandis que Pavlov arrive premier au niveau du béhaviorisme (0.19). Enfin, Vallerand n'est pas associé avec la psychologie sociale, contrairement à la réalité. Il est possible d'expliquer ceci en référant aux caractéristiques des mots employés. En effet, les mots psychologie et sociale sont utilisés dans bon nombre de contextes et indépendamment de surcroît, sans doute sans lien précis avec les citations du professeur Vallerand.

En définitive, il est important de garder à l'esprit la nature statistique de la LSA. À cet égard, on peut penser que les analyses opérées à partir de documents plus longs, tels que ceux utilisés dans la présente étude, sont moins sujettes aux aberrations découlant d'une analyse opérées sur un ou deux mots.

Exemple de système de tutorat intelligent : Le Macro-Professeur

La section suivante se divise en trois parties. En premier lieu, une brève description du logiciel ainsi que les principales étapes de son élaboration sont exposées. En deuxième lieu, un aperçu de l'interface du logiciel est présenté. Enfin, la dernière partie décrit la démarche d'ajustement de certains paramètres du logiciel.

Description et démarche d'élaboration du logiciel Le Macro-Professeur

Le logiciel *Le Macro-Professeur* est disponible gracieusement sur le site Internet de la LSA à l'adresse suivante : http://lsa.colorado.edu/LeMacro-Professeur/. *Le Macro-Professeur* oriente l'utilisateur vers la production d'un résumé comportant l'information essentielle (2) d'un texte tout en satisfaisant une contrainte de longueur du résumé. Sa mécanique fait appel *à* la *Latent Semantic Analysis*. L'espace sémantique, décrit précédemment, constitue la base de connaissances permettant au logiciel d'évaluer les résumés de textes prédéterminés et de fournir de la rétroaction à l'utilisateur. L'interface par lequel l'utilisateur prend contact avec les stimuli et la rétroaction du logiciel est composé de pages HTML. Le logiciel ne requiert donc pour son utilisation qu'une plateforme *PC* ou *Mac* munie du logiciel de navigation *Microsoft Internet Explorer* 5 ainsi que d'un accès à Internet. Le logiciel *Microsoft Virtual Machine* est téléchargé automatiquement et gratuitement lors de l'utilisation du logiciel s'il n'est pas déjà présent dans l'ordinateur. Les étapes suivantes ont été nécessaires à la réalisation du *Macro-Profeseur*.

D'abord, tous les messages de rétroaction de la version anglaise ont été traduits en français. Il s'agit d'une traduction libre de l'auteur principal. Ces messages ont ensuite été intégrés à l'interface du logiciel afin de remplacer les

(2) La détermination de l'information essentielle a un texte est discutée ultérieurement.

messages originaux en anglais. Il est à noter que cette procédure ne requiert nullement de dispositions particulières comme dans le cas de la validation transculturelle d'un questionnaire. Il s'agit de faire en sorte que les messages de rétroaction du logiciel soient compréhensibles par l'utilisateur francophone. Comme les processus de compréhension en lecture de même que l'habileté de résumer sont des mécanismes cognitifs qui n'impliquent pas à prime abord de composantes culturelles, le logiciel présenté dans cette étude est destiné à toute personne qui maîtrise le français écrit.

Deuxièmement, trois textes servant d'exercices, choisis en raison de leur intérêt général, ont été intégrés au logiciel. Les textes sont présentés en ordre croissant de difficulté, sur la base d'une expérimentation sommaire. Ces trois textes sont tirés de trois des six livres ayant servi à l'élaboration de l'espace sémantique, afin d'assurer la plus grande correspondance entre les textes à traiter par le logiciel et l'espace sémantique. Les titres des textes sont les suivants: *L'expérience des études collégiales et universitaires, Le modèle de l'intelligence artificielle (IA)* et *L'amour.*

Le Macro-Professeur est développé en vue d'un usage très spécifique : aider les étudiants de niveau universitaire à développer leur habileté à résumer des textes de vulgarisation scientifique. En raison des limites reliées à l'intelligence artificielle à laquelle il fait appel, le logiciel propose présentement du matériel orienté vers les sciences humaines.

L' interface du logiciel

L'interface du logiciel *Le Macro-Professeur* est à plusieurs égards comparable au matériel courant disponible sur l'Internet. En fait, l'environnement interactif du logiciel est composé de fenêtres de saisie pour le texte et de boutons d'action proposant les différentes options. Ces éléments, de même que la rétroaction, sont intégrés et présentés dans une succession de pages web spécifiques au logiciel et organisées autour de la mécanique sous-jacente au logiciel, la LSA. Cette approche permet un interface attrayant, avec logos et motifs d'arrière-plan, que de brèves instructions fournies sur la page d'accueil permettent de rendre convivial.

La page d'accueil du logiciel est élaborée en vue d'une utilisation autonome, notamment en raison de son accessibilité via Internet. Ceci afin que quiconque entre en contact avec le site, soit par hasard ou en y étant référé, puisse y évoluer avec succès sans instructions supplémentaires. Après une brève description de la nature d'un résumé et l'énumération des stratégies de production d'un bon résumé (3), la page contient les instructions détaillées au niveau de l'utilisation du logiciel. Enfin, cette page se termine par des informations concernant l'utilisation de la rétroaction fournie par *Le Macro-Professeur.*

En plus de la page d'introduction et de pages connexes proposant des informations supplémentaires, l'interface du logiciel est composé d'une fenêtre de saisie permettant à l'utilisateur de taper son résumé au moyen du clavier de l'ordinateur. De plus, cette fenêtre indique la longueur idéale du résumé à produire. Cette longueur idéale est d'environ 7.6 % de la longueur du texte original, ce qui contribue à rendre le niveau de difficulté des exercices proposés plutôt élevé. Mis à part la rédaction du résumé réalisée au moyen du clavier, l'interaction avec le logiciel se fait au moyen de la souris. L'utilisateur l'utilise pour cliquer sur les hyperliens afin d'obtenir plus d'information. Il doit aussi cliquer sur les différents boutons d'action qui proposent les choix d'action possibles en fonction de la progression de l'utilisateur dans la tâche à réaliser. Après la rédaction du résumé, Le Macro-Professeur propose notamment de sauvegarder le résumé produit et d'en vérifier l'orthographe ou de fournir la rétroaction sur ce résumé.

La page de rétroaction comporte une partie graphique suivie d'une partie verbale. On y retrouve un indicateur de la longueur du résumé, représenté par la bande verticale. Si l'extrémité de cette bande se situe entre les deux lignes vers le milieu de l'indicateur, la longueur du résumé est adéquate, (tel que symbolisé par la couleur verte de l'indicateur.) Si l'extrémité de la bande se situe en deçà ou au-delà des limites, l'indicateur passe au rouge et signifie que le résumé est trop court ou trop long. En soi, un résumé ne peut être trop court, sauf si sa longueur ne permet pas de couvrir adéquatement l'information principale du texte à résumer. À cet égard, les limites inférieures sont calibrées de façon à ce qu'un résumé très court oriente la rétroaction à l'effet que le contenu du texte n'est pas couvert possiblement en raison de sa brièveté. Les bandes horizontales indiquent la quantité d'information contenue dans le résumé. Chaque bande correspond à une section du texte à résumer et est identifiée par le titre de la section. La ligne noire verticale représente le seuil à atteindre (4) pour chaque section. Ce seuil symbolise la quantité minimale d'information que chaque section du résumé doit contenir pour représenter l'information principale du texte.

En complément, la page de rétroaction comporte certains commentaires verbaux. Ces commentaires fournissent des hyperliens et des boutons d'action adaptés à la qualité du résumé de l'usager, de même que des renseignements complémentaires aux graphiques. Il traitent de la qualité du résumé et proposent des stratégies afin de l'améliorer.

Quand son résumé rencontre les normes de longueur et de contenu, le logiciel lui offre de vérifier si deux ou plusieurs phrases du résumé contiennent la même information (redondance) ou si chaque phrase est bel et bien en

(3) Ces stratégies sont énoncées d'après Brown et Day (1983).

(4) La méthode de détermination de ce seuil est discutée plus loin.

lien avec le sujet du texte (pertinence).

Enfin, de par la relative simplicité de l'environnement du logiciel, les possibilités d'amélioration sont nombreuses. Il est possible d'intégrer une quantité infinie de textes servant d'exercices de difficulté variée afin d'adapter Le *Macro-Professeur* à toute la clientèle estudiantine, du primaire au doctorat. Il est aussi possible d'élargir les domaines couverts par les exercices, de la chimie à la musicologie, moyennant la création de l'espace sémantique approprié.

Ajustement des paramètres du logiciel

Cette section décrit deux principaux paramètres du logiciel. Le premier paramètre est le seuil d'importance du contenu. Le deuxième paramètre est la longueur du résumé à produire. La section se termine par l'appréciation du degré de difficulté du logiciel.

Démarche automatisée d'identifcation du contenu le plus important d' un text.

Au sein de cette méthode, le jugement humain n'intervient jamais dans la détermination du contenu important du texte à résumer. Les étapes mathématiques qui suivent décrivent l'essentiel de cette démarche.

1. Diviser le texte en sections S 1.. . Sn.
2. Pour chaque section Si du texte, la diviser en phrases sl, s2, . . . sm.
3. Opérer les comparaisons C 1 = cos(s1,Si), C2 = cos (s2,Si), . . . Cm = cos (sm, Si).
4. Choisir la phrase sj telle que cos(sj,Si) = max(C1.. . Cm).
5. Prendre les n phrases obtenues et les réunir pour former un résumé.
6. Seuil i = cos(résumé, Si).

En somme, le système décompose le texte original en sections selon les sous-titres. Ensuite, il identifie les phrases composant chaque section. Il produit un indice de similarité sémantique entre chaque phrase et la section d'où est tirée la phrase. Il choisit ensuite la phrase qui affiche la plus grande similarité avec sa section. On joint ces phrases les plus représentatives de leur section pour obtenir un «résumé typique». La dernière étape sert à déterminer le seuil à atteindre pour chaque section du texte à résumer, c'est-à-dire le contenu minimal qui doit être présent dans le résumé de l'utilisateur du *Professeur.*

Cette démarche est déterminante pour l'efficacité du logiciel. En effet, la

mécanique du Macro-Professeur oriente l'utilisateur vers la production d'un résumé dont le degré de similarité sémantique avec le texte original excède celui du contenu déterminé par la présente démarche. C'est en quelque sorte un seuil qui détermine artificiellement la proportion du texte à résumer qui constitue sa macrostructure. Plus ce seuil est élevé, plus petite est la portion du texte qui renferme sa macrostructure. Par conséquent, la partie du texte original à négliger dans le résumé prend de l'ampleur. Il est possible qu'un seuil plus élevé favorise l'utilisation de stratégies de haut niveau dans l'élaboration du résumé.

La contrainte de longueur du résumé à produire

La contrainte de longueur est fixée empiriquement dans le cadre de cette étude à environ 7.6% de la longueur du texte original. Par exemple, le sujet qui choisit le texte portant sur l'intelligence artificielle (2392 mots) doit produire un résumé comportant entre 250 et 300 mots.

Degré de difficulté du logiciel

Le seuil symbolisant le contenu minimal que le résumé doit contenir ainsi que la contrainte de longueur du résumé constituent les deux paramètres permettant de faire varier la difficulté de la tâche. Le seuil déterminé par la méthode décrite précédemment impose de ne retenir que l'information la plus importante du texte à résumer. La contrainte de longueur, quant à elle, impose la formulation concise et précise du contenu important du texte à résumer. L'ajustement actuel de ces paramètres, l'un automatiquement et l'autre empiriquement, font de la tâche proposée par *Le Macro-Professeur* un défi exigeant pour les étudiants universitaires, qui met à l'épreuve leurs capacités de compréhension en lecture.

Références

Aamoutse, C. A. J., van den Bos, K. P., & Brand-Gruwel, S. (1998). Effects of listening comprehension training on listening and reading. *The Journal of Special Education, 32* (2), 115-126.

Aaronson D. (1994). Computer use in cognitive psychology. *Behavior Research Methods, Instruments, & Computers, 26,* (2), 81-93.

Assemblée des gouverneurs. (1999). *Les études de premier cycle : politiques et règlements* (5e éd.) [Brochure]. Université du Québec.

Brown, A.L., & Day, J. D. (1983). Macrorules for summarizing texts : The development of expertise. *Journal of Learning and Verbal Behavior, 22,* 1-14.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space : Words, sentences, discourse. *Discourse Processes,* 2.5 (2 & 21 l-257.

Foltz, P.W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28* (2), 197-202.

Frederiksen, C.H., & Donin, J. (1999). *Cognitive Assessment in Coached Learning Environments.* Manuscrit soumis pour publication.

Kintsch, W. (1998). *Comprehension : A paradigm for cognition.* New-York : Cambridge University Press.

Landauer, T.K. & Dumais, S.T. (1996). How corne you know so much? From practical problems to new memory theory. Dans Hermann, D.J., McEvoy, C., Hertzog, C., Hertel, P., & Johnson, M.K. (Éds), *Basic and applied memory research : Vol. 1. Theory in context* (pp.105-126). Mahwah, N.J. : Lawrence Erlbaum Associates, Inc.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem : The Latent Semantic Analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review,* 104,21 l-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes,* 25,259-284.

Landauer, T.K., Laham, D., Rehder, B., & Schreiner, M.E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. Dans Shafto, M.G. & Langley (Éds), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ : Erlbaum.

Lillehaug, S.-I., & Lajoie, S.P. (1998). AI in medical education - another grand challenge for medical informatics. *Artificial Intelligence in Medicine,* 12,197-225.

McNamara, D.S., Kintsch, E., Songer, B.N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction,* 14 (l), l-43.

Mercier, J. (1999). *Espace sémantique Psychology_French,* http://lsa.colorado.edu/.

Perfetti, C.A. (1998). The limits of co-occurrence : Tools and theories in language research. *Discourse Processes, 25* (2&3), 363-377.

Rehder, B., Schreiner, M.E., Wolfe, M.B.W., Laham, D., Landauer T.K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge : Some technical considerations. *Discourse Processes, 25* (2&3), 337-354.

Les Auteurs

Julien Mercier est une étudiante au doctorat à la Université

Monique Brodeur est une docteur a la Université du Québec à Montréal.