

Dessus, P., Lemaire, B. (1999). Sciences et Techniques Éducatives, 6-2, 409-415.

APex, un système d'aide à la préparation d'examens

1. Introduction

Les avancées de l'intelligence artificielle parviennent parfois jusqu'au grand public, pour peu qu'une « mutation » visible, avec une dimension sociale, puisse être mise en avant. Par exemple, la disparition d'un corps de métier — ici les enseignants — au profit de machines. C'est ce qui s'est passé en 1998 avec la conception de l'*Intelligent Essay Assessor* (IEA) de Foltz, Laham et Landauer [FOL 99], logiciel qui permet de noter des dissertations par le biais d'un appariement sémantique de ces dernières avec des copies modèles sélectionnées au préalable par l'enseignant. Un débat médiatique s'est organisé, tout d'abord américain [HOL 98], [KAH 98], [PER 98], puis français [ZIL 98], sur le thème désormais classique : « Les ordinateurs mettront-ils un jour les enseignants à la porte ? » [LAM 98].

Tout en essayant de ne pas verser de nouvelle polémique dans ce débat, nous allons ici présenter les fonctionnalités d'IEA ainsi que celles du logiciel que nous élaborons, APex, qui utilise le même moteur que le premier mais procède de manière sensiblement différente. Commençons par décrire brièvement le moteur commun de ces deux logiciels : LSA pour *Latent Semantic Analysis* (analyse sémantique latente).

2. LSA, une méthode d'analyse factorielle multidimensionnelle

LSA [LAN 97]¹ est un modèle statistique fondé sur un type d'analyse factorielle permettant d'analyser la proximité sémantique intermots ou paragraphes à l'intérieur d'un grand ensemble d'unités d'informations textuelles. Initialement conçu pour améliorer l'efficacité de l'interrogation de systèmes documentaires informatisés, le modèle de LSA suppose que, étant donné plusieurs « contextes » (unités d'information textuelle, soit phrases, paragraphes, discours...), il existe une structure latente dans l'usage des mots communs à ces contextes et qu'une analyse statistique permet de mettre en évidence cette structure.

Le modèle de LSA pose que la similarité sémantique de deux mots est liée à la probabilité que ces mots se retrouvent dans le même contexte, ou dans deux contextes similaires. Toutefois, LSA ne tient compte que de l'appartenance des mots au contexte, et non de leur ordre au sein des phrases. Il peut donc considérer comme proches deux mots n'apparaissant jamais dans le même contexte, mais dont les contextes respectifs contiennent des mots similaires. Il est important de noter que cet

¹ LSA est écrit en langage C et fonctionne sur une station de travail Unix, il est déposé en 1990 par *Bell Communications Research Inc.* Voir <<http://lsa.colorado.edu>> pour de nombreuses informations sur LSA.

appariement est d'autant plus juste que le corpus de textes traités est important. LSA permet de calculer la proximité sémantique entre deux termes (ou un terme et un contexte, ou encore deux contextes) et donne un indice d'autant plus élevé que ces deux entités sont de sens voisin ou bien ont été fréquemment associées. Ces relations que LSA détecte entre les mots résultent d'une réduction de dimensions de la matrice d'occurrences des mots dans les différents contextes, par le biais d'une décomposition aux valeurs singulières [DEE 90].

Passons maintenant à la description des deux logiciels utilisant LSA pour appairer les productions d'étudiants à des copies ou textes de référence, IEA et APex, pour *Assistance à la préparation des examens*.

3. IEA et APex, des aides à la correction automatique de copies

De nombreux chercheurs en éducation et en psychologie se sont récemment intéressés aux potentialités de LSA. Ils l'ont notamment testé dans les domaines suivants, en mettant au jour des performances de LSA comparables avec des performances humaines :

- modélisation de l'apprentissage [LAN 97], [LEM 98] ;
- mesure de la cohérence textuelle [FOL 98] ;
- tuteurs intelligents [LEM 99], [WIE 98] ;
- génération de liens hypertextes [DES 99] ;
- analyse de communications dans un campus virtuel [NOL 98] ;
- correction automatique de résumés [PAU 99] ou de copies [FOL 99].

Intéressons-nous de plus près à ce dernier point.

3.1. IEA, *Intelligent Essay Assessor*

Lorsque un enseignant veut s'assurer que ses élèves ont assimilé son cours, il peut réaliser des questionnaires à choix multiple qui seront remplis par les étudiants et corrigés, automatiquement ou non. Une autre possibilité est de demander à l'élève de rédiger un texte correspondant à ce qu'il connaît d'un cours et, par une méthode adéquate, de comparer ce texte à des textes-cibles préalablement sélectionnés. C'est ce que réalise IEA : après avoir « entraîné » LSA avec un corpus du domaine (cours), le texte de l'élève est comparé à une ou plusieurs copies-types, sélectionnée(s) par l'enseignant. Foltz *et al.* [FOL 99] ont testé deux techniques, donnant deux types de scores à la copie :

— le score « holistique », qui compare successivement le texte à noter à une série de copies notées au préalable par un jury. La note de la copie sera celle de la série de copies avec laquelle elle entretient la plus grande proximité, calculée par IEA. Une évaluation de ce calcul de score a été faite à partir de 190 copies de biologie, elle montre une corrélation de .80 entre les scores des évaluateurs humains et ceux calculés par IEA.

— le score « étalon-or » (*gold standard*), qui compare le texte à noter avec une copie-modèle idéale, réalisée par exemple par l'enseignant. La comparaison peut être

globale ou bien faite paragraphe par paragraphe, de manière à vérifier si l'élève traite correctement chaque notion.

Un des avantages majeurs d'IEA est que l'étudiant peut soumettre son texte autant de fois que nécessaire, tant que ce dernier n'obtient pas la note voulue. Foltz *et al.* observent d'ailleurs que les notes moyennes des textes soumis à IEA passent de 85/100 — note de la première soumission — à 92/100.

3.2. APex, une Assistance à la préparation des examens

APex, tout en utilisant le même moteur, LSA, procède différemment, en ce que l'enseignant n'a pas besoin de sélectionner des copies-types. La copie de l'étudiant est ainsi comparée directement au cours de l'enseignant². Un corpus de textes en français est traité conjointement à cette copie, de manière à ajouter des connaissances de la langue, qui augmentent ainsi la précision des comparaisons sémantiques. L'enseignant peut empiriquement régler les seuils d'évaluation amenant un commentaire (notion très bien, bien, mal ou très mal relatée). Enfin, les visées d'APex ne sont pas la production d'une note, mais plutôt des évaluations plus fines sur les trois niveaux suivants, de finesse croissante :

— *au niveau du contenu*, en appariant la copie avec chacune des notions du thème choisi, ce qui permet d'évaluer la manière dont le contenu a été couvert ;

— *au niveau du plan*, en appariant chaque paragraphe de la copie avec chaque notion du thème choisi, ce qui permet à l'étudiant d'appréhender le plan général de sa copie ;

— *au niveau de la cohérence textuelle*, en mesurant successivement la proximité sémantique de deux phrases contiguës de la copie, ce qui permet d'alerter l'étudiant si la microcohérence de sa copie est insuffisante, en raison de ruptures brutales de cohérence interphrases.

L'étudiant peut ainsi, à tout moment, obtenir une évaluation de sa copie à propos de l'un ou l'autre de ces trois niveaux (*voir annexe*), ce qui autorise une planification et une révision de sa copie plus fréquentes et précises.

Le premier test que nous avons réalisé porte sur un cours de maîtrise de sciences de l'éducation (ergonomie de la formation). Après avoir fait traiter le cours par APex, nous avons récupéré cinq copies d'examen. Nous trouvons, au niveau du contenu, une corrélation moyenne de .51 ($p < .01$) entre les notes partielles de l'enseignant, données pour évaluer le traitement par l'étudiant de chacune des huit notions requises et celles données par APex. Le niveau du plan a également été testé et donne des résultats satisfaisants, bien que devant être faits sur un nombre plus important de copies.

4. Travaux à venir

² Cela suppose que l'enseignant réalise un découpage de son cours en deux niveaux de hiérarchie : les thèmes principaux et les notions qui se rattachent à chaque thème. Un système de balises permet à APex de récupérer cette structure pour la comparer avec les copies qu'on lui soumet.

APex doit faire l'objet de tests plus approfondis — au niveau des corpus de connaissance de la langue et au niveau du réglage des seuils d'évaluation. Dans des travaux à venir nous allons également incorporer un modèle de l'élève qui sera traité par LSA, ce qui permettra de proposer à l'étudiant des sujets plus conformes à ce qu'il connaît. Des modules de calcul de lisibilité et de génération de liens hypertextes sont également en cours de réalisation. Des études plus spécifiquement didactiques nous permettront de vérifier la compatibilité d'APex avec les modèles de production de textes en vigueur [BER 87], et d'en proposer une version interrogeable à distance.

L'intérêt d'APex réside en ce qu'il dispense l'enseignant de réaliser une analyse du contenu profonde, donc complexe (par exemple en termes de réseaux sémantiques, de règles de logique, de scripts). Ici, seuls suffisent une structuration du cours en deux niveaux de hiérarchie, un choix adéquat de la base textuelle modélisant des connaissances de la langue³, ainsi qu'un réglage des seuils d'évaluation.

*

Plutôt que de préconiser prématurément une reconversion massive des enseignants en des métiers non encore touchés par l'informatisation, nous préférons développer avec APex des situations où l'enseignant serait libéré de certaines tâches fastidieuses et où les élèves bénéficieraient d'évaluations plus fréquentes et contextualisées.

Philippe DESSUS & Benoît LEMAIRE
Lab. des Sciences de l'éducation
Université de Grenoble
{Prenom.Nom}@upmf-grenoble.fr

5. Références bibliographiques

- [BER 87] BEREITER, C., SCARDAMALIA, M., *The psychology of written composition*, Erlbaum, Hillsdale, 1987.
- [DEE 90] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., HARSHMAN, R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, 1990, p. 391-407.
- [DES 99] DESSUS, P., « Vérification sémantique de liens hypertextes », *Conférence Hypertextes et Hypermédias : Réalisations, Outils & Méthodes*, Univ. de Paris VIII, Paris, 23-24 sept., 1999.

³ Nous projetons également d'effectuer un travail plus conséquent sur cette base textuelle, qui joue un rôle important dans la manière qu'a LSA d'appréhender la sémantique des termes qu'il traite. Pour l'instant, des tests limités nous ont fait choisir trois romans du XIX^e siècle, comprenant 290 000 mots. L'utilisation d'un corpus plus important et adéquat est à l'étude.

APex, un système d'aide à la révision d'examens 5

- [FOL 98] FOLTZ, P. W., KINTSCH, W., LANDAUER, T. K., « The measurement of textual coherence with Latent Semantic Analysis », *Discourse Processes*, vol. 25, n° 2-3, 1998, p. 285-307.
- [FOL 99] FOLTZ, P. W., LAHAM, D., LANDAUER, T. K., « Automated essay scoring : applications to Educational Technology », *Proc. ED-MEDIA '99*, Seattle, 1999.
- [HOL 98] HOLMES, B., « Mark my words... student essays can now be graded by machines », *New Scientist*, n° 2131, 1998, p. 12.
- [KAH 98] KAHNEY, L., « Teachers of Tomorrow? », *WIRED News Online*, 3 nov., 1998.
- [LAM 98] LAMBERT, E., « Will Computer give Professors the Boot? », *The Colorado Daily*, 16 avril, 1998.
- [LAN 97] LANDAUER, T. K., DUMAIS, S. T., « A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge », *Psychological Review*, vol. 104, 1997, p. 211-240.
- [LEM 98] LEMAIRE, B., « Models of High-dimensional Semantic Spaces », *Proc. 4th Int. Workshop on Multistrategy Learning (MSL '98)*, Desenzano, juin, 1998.
- [LEM 99] LEMAIRE, B., « Tutoring Systems based on Latent Semantic Analysis », *Proc. 9th Int. Conf. on Artificial Intelligence in Education (AI-ED '99)*, Le Mans, 19-23 juillet, 1999.
- [NOL 98] NOLAN, J., *Educators in MOOkti : a polysynchronous Collaborative Virtual Learning Environment*, Thèse de l'Ontario Institute for Studies in Education, Toronto, 1998.
- [PAU 99] PAULA, R. de, *Summary street*, <<http://lsa.colorado.edu/summarystreet/>>, 1999.
- [PER 98] PERLSTEIN, L., « New Software checks Essay Content », *The Washington Post*, 13 Oct., 1998 (trad. fr. dans *Courrier International*, n° 418, 5-11 nov., 1998, p. 48).
- [WIE 98] WIEMER-HASTING, P., GRAESSER, A. C., HARTER, D., « The foundations and Architecture of Autotutor », In B. P. Goettl, H. M. Half, C. L. Redfield, V. J. Shute (Eds), *Intelligent Tutoring Systems (ITS '98)*, Springer Verlag, Berlin, 1998, 334-343.
- [ZIL 98] ZILBERTIN, O., « Un logiciel pour corriger les copies des étudiants américains », *Le Monde*, 12 nov., 1998.

Annexe

Exemples d'analyse d'une copie d'étudiant par APex dans les trois niveaux : contenu, plan, cohérence. Les valeurs sont comprises entre -1 et 1.

Niveau du contenu :

Analyse du document : Longueur du document : 658 mots. NOTE GENERALE : 9.7 / 20

6 Sciences et techniques éducatives. Volume x – n° x/199x

Vous avez très mal relaté :

- Erreurs basées sur les automatismes (ratés et lapsus) (0.26)
- Une étude à propos de la résolution de problèmes (Testu & B... (0.25)

Vous avez mal relaté :

- Erreurs basées sur les connaissances déclaratives (fautes) (0.40)
- Le modèle de Testu (1993) processus contrôlés vs automatisé... (0.34)

Vous avez bien relaté :

- Définition de l'activité (0.59)
- Le modèle de l'activité de Rasmussen (0.70)
- Erreurs basées sur les règles (fautes) (0.57)

Vous avez très bien relaté :

- Transposition du modèle de l'activité de Rasmussen à l'ense... (0.77)

Niveau du plan :

Paragraphe 1 : Avant de décrire le modèle de Rasmussen rappelo...
Notion la plus proche (0.72) : Définition de l'activité

Paragraphe 2 : Au niveau 1 Rasmussen place le comportement ba...
Notion la plus proche (0.77) : Transposition du modèle de l'activité

Paragraphe 3 : La hiérarchie de niveaux s'explique dans le sens ...
Notion la plus proche (0.65) : Transposition du modèle de l'activité

Paragraphe 4 : Prenons une application précise du modèle de Ras...
Notion la plus proche (0.91) : Transposition du modèle de l'activité

Paragraphe 5 : Reason a établi un catalogue d'erreurs possibles ...
Notion la plus proche (0.60) : Le modèle de l'activité de Rasmussen

Paragraphe 6 : Les erreurs d'attention excessive proviennent du f...
Notion la plus proche (0.75) : Erreurs basées sur les automatismes

Paragraphe 7 : Au niveau du comportement basé sur les règles l...
Notion la plus proche (0.90) : Erreurs basées sur les règles (fautes)

Paragraphe 8 : Au niveau du comportement basé sur les connaissanc...
Notion la plus proche (0.75) : Erreurs basées sur les connaissances dé

Niveau de la cohérence (extraits) :

- 1: Avant de décrire le modèle de Rasmussen rappelons la défi...
- 2: L'activité est un ensemble de comportements inobservables l...
- 3: On distingue généralement deux grandes catégories d'activ...
- 4: Rasmussen lui décompose l'activité en trois types de compo...
- 5: Si l'on parle d'activité en échelle de Rasmussen c'est que...
- 6: Au niveau 1 Rasmussen place le comportement basé sur les ha...

coherence phrase 1 - phrase 2 : 0.67

coherence phrase 2 - phrase 3 : 0.40

coherence phrase 3 - phrase 4 : 0.58

coherence phrase 4 - phrase 5 : 0.79

coherence phrase 5 - phrase 6 : 0.52

coherence inter-phrases médiocre (0.49)