

Acquisition de connaissances à partir de textes : modèles cognitif et computationnel

Virginie Zampa et Benoît Lemaire
L.S.E.

Université Grenoble 2
BP 47 - 38040 Grenoble Cedex 9

Virginie.Zampa@upmf-grenoble.fr, Benoit.Lemaire@upmf-grenoble.fr

textes. En effet, deux mots peuvent être considérés

1 Introduction

Nous disposons d'un modèle *cognitif* d'acquisition des connaissances à partir de textes qui possède la particularité d'être implémenté, et nous nous proposons d'étudier les possibilités d'application de ce modèle à l'acquisition et à la représentation *informatique* de connaissances. Les représentations obtenues ouvrent de nombreuses perspectives (1) pour la modélisation du domaine et de l'élève dans le cadre d'un tuteur intelligent ; (2) pour l'aide à la construction et à la validation d'ontologies.

2 Le modèle cognitif LSA

Des travaux en psychologie cognitive montrent que la majorité des mots sont acquis par la lecture [4]. Etant exposé à des textes, l'apprenant va petit à petit affiner le sens des mots grâce aux occurrences conjointes de ces mots avec d'autres. Par exemple, sans le lui définir explicitement, l'apprenant va acquérir le sens du mot "abat-jour" parce que ce mot apparaît avec d'autres comme "ampoule", "éclairer", "lumière" dans les textes qu'il lit. Or, il semble que ce n'est pas simplement la cooccurrence répétée d'un mot avec d'autres qui permet l'acquisition du sens du mot, mais plutôt l'ensemble des cooccurrences de tous les mots au fil des textes. Landauer et Dumais [4] ont conçu sur cette base un modèle de représentation des connaissances : l'analyse de la sémantique latente (LSA).

Ce modèle cognitif a été implémenté à l'aide de l'outil LSI, conçu au début des années 90 pour la recherche sémantique d'informations dans les textes [1]. Il s'appuie sur une représentation multidimensionnelle des mots de la langue. Grâce à une analyse statistique de grands corpus, le sens de chaque mot est caractérisé par un vecteur dans un espace de grandes dimensions, la proximité entre deux vecteurs correspondant à la proximité de sens de ces mots.

Cette analyse statistique consiste à construire une matrice d'occurrences qui sera réduite (par une procédure que nous présenterons plus loin) afin de faire ressortir les relations sémantiques entre mots ou entre

sémantiquement proches s'ils sont utilisés dans des contextes similaires. Le contexte d'un mot est ici défini comme l'ensemble des mots qui apparaissent conjointement avec lui. Ainsi, les mots *vélo* et *bicyclette* sont considérés sémantiquement proches car ils apparaissent tous deux avec des mots comme *randonnée*, *guidon*, *pédaler*, etc. et ils n'apparaissent pas avec des mots comme *bouillir*, *imprimante*, *canard*, etc. Cette notion de cooccurrence est évidemment statistique : la méthode fonctionne si un nombre suffisant de textes est utilisé.

La procédure utilisée est la réduction de la matrice d'occurrences des mots dans les paragraphes, non pas jusqu'à 2 ou 3 comme en analyse factorielle classique, mais jusqu'à une centaine de dimensions. Ce nombre est important car une réduction à un espace de trop grande dimension ne ferait pas suffisamment émerger les liaisons sémantiques entre mots, et un trop petit nombre de dimensions conduirait à une trop grande perte d'informations. Le nombre adéquat de dimensions ne peut pas être actuellement déterminé théoriquement ; seuls des tests empiriques ont permis de situer cette valeur entre 100 et 300 dans le cas de l'anglais [2].

L'espace sémantique étant construit, la proximité sémantique entre deux mots est déterminée par le cosinus de leur angle.

Les associations sémantiques extraites par LSA ne se réduisent pas uniquement à des relations de synonymie, puisqu'elles proviennent d'un traitement de la cooccurrence des termes avec leurs contextes. Ces associations sémantiques ont l'avantage d'être relativement générales mais possèdent l'inconvénient de ne pouvoir être caractérisées plus finement.

Plusieurs travaux ont été entrepris pour valider cette représentation sémantique multidimensionnelle. A partir d'un corpus de 4,6 millions de mots tirés d'une encyclopédie, LSA a obtenu sur la partie « synonymes » du test du TOEFL, un score comparable à celui des sujets non anglophones postulant à l'entrée dans les universités américaines (64,5%). Aucun système n'est aujourd'hui capable d'avoir un tel score en se fondant uniquement sur l'analyse automatique de textes, c'est-à-dire sans aucun recours à un codage manuel de connaissances.

Le fait que ce modèle cognitif soit opérationnel rend envisageable son application en intelligence artificielle.

Par rapport à des représentations sémantiques de type symboliques (réseaux sémantiques, réseaux terminologiques), on peut noter que la représentation de LSA a le désavantage de n'avoir qu'un seul type de lien. En revanche, la méthode est entièrement automatique et indépendante du domaine. Elle est donc applicable à des corpus très volumineux.

3 Représentation des connaissances dans un tuteur intelligent

Il s'agit maintenant de représenter les connaissances sur le domaine et sur l'apprenant. Les connaissances sur le domaine résultent de l'analyse par LSA de textes experts. Les connaissances sur l'apprenant proviennent de l'analyse de textes produits par l'apprenant. C'est la confrontation de ces deux modèles de représentation de connaissances qui va permettre de diagnostiquer des erreurs de l'apprenant (mots voisins différents dans les deux espaces) et de sélectionner les meilleurs textes à faire lire à l'apprenant pour que son apprentissage soit maximal [5,7].

4 Application à la construction d'ontologies

LSA peut être un élément de réponse à la question de la réutilisabilité des ontologies. Par exemple, il est possible d'imaginer une base ontologique dans le domaine financier, créée par LSA à partir de l'analyse de grands corpus du domaine. Cette base contiendrait alors tous les termes financiers, représentés dans l'espace sémantique. La construction d'une ontologie pour une application particulière ne se réaliserait pas *ex nihilo*, mais en s'appuyant sur la base. Il s'agirait alors à l'expert de spécifier les relations entre les termes que LSA a jugé proches. Il est aisé d'imaginer une interface où LSA proposerait un réseau terminologique vierge que l'expert étiquetterait petit à petit. Il est également possible de s'appuyer sur une base ontologique générale, issue de l'analyse d'une encyclopédie entière. En effet, des travaux ont montré l'intérêt de disposer de ressources générales dans la construction d'ontologies spécialisées [3].

Dans la même lignée, LSA pourrait être utilisée pour la validation d'ontologies, en détectant des associations entre des termes éloignés dans l'espace. L'expert, là encore, serait seul juge de la confirmation des suggestions de la machine. LSA peut donc, se révéler un complément intéressant à des méthodes davantage dépendantes d'un domaine.

5 Conclusion

En dépit d'une représentation moins riche que les autres méthodes d'extraction de connaissances, LSA possède l'avantage d'être :

- indépendante du domaine et complètement automatique
- validée expérimentalement sur de nombreux domaines.

Elle nous paraît prometteuse pour la construction ou la validation d'ontologies, en complémentarité avec des méthodes dépendantes du domaine. LSA est en effet bien plus qu'une méthode de traitement statistique de textes : Rehder et al. [6] ont montré que LSA permet de mesurer la connaissance d'un sujet sur un domaine à partir de l'analyse de ses textes, de manière comparable à des juges humains. C'est cette possibilité de représenter de la connaissance qui en fait un outil intéressant pour notre propos.

L'absence de tout recours à la syntaxe reste une caractéristique étonnante de LSA : le sens d'une phrase peut être analysé avec succès (en comparaison avec les performances des humains) sans que l'ordre de ses mots ne joue un rôle. C'est là un autre aspect intéressant de LSA que de s'affranchir de ce niveau syntaxique par trop difficile à modéliser.

6 Références

- [1] S.T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, vol 23(2), pages 229-236, 1991.
- [2] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer et R. Harshmann. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, vol 41, pages 391-407, 1990.
- [3] T. Hamon, D Garcia et A. Nasarenko. Détection de liens de synonymie : complémentarité des ressources générales et spécialisées. *Actes de la conférence "Terminologie et IA."*, Nantes page 45-58, mai 1999.
- [4] T.K. Landauer et S.T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review* vol 104(2), pages 211-240, 1997.
- [5] B. Lemaire. Tutoring systems based on Latent Semantic Analysis. In S.P. Lajoie, M. Vivet (Eds) *Artificial Intelligence in Education (proceedings of the AIED'99 Conference)*, pages 527-534, IOS Press, 1999.
- [6] B. Rehder, M.E. Schreiner, M.B. Wolfe, D. Laham, T.K. Landauer, et W. Kintsch. Using Latent Semantic Analysis to assess knowledge: Some technical considerations, *Discourse Processes*, vol 25, pages 337-354, 1998.
- [7] V. Zampa, Automatic Texte Selection by Latent Semantic Analysis. *Proceedings of the Young Researchers Track*In S.P. Lajoie, M. Vivet (Eds)

Artificial Intelligence in Education (proceedings of the AIED'99 Conference), pages 535-542, IOS Press, 1999.