Traduisez ce texte 1 en Arabe

Aperçu des techniques mises en œuvre

Il n'y a pas lieu de décrire ici en détail les techniques mises en œuvre. On pourra toutefois mettre en avant deux ou trois grands types d'approches :

Dans les années 1980, la plupart des systèmes visent une analyse approfondie du contenu et, de fait, mettent en jeu des systèmes de connaissances et de représentation très fouillés. Ces systèmes sont par conséquent très coûteux à mettre en œuvre et peu portables d'un domaine à l'autre.

Les années 1990 voient fleurir les systèmes fondés sur la technologie à nombre fini d'états (automates et/ou transducteurs à nombre fini d'états). Des résultats théoriques avaient démontré que cette technologie n'était pas suffisante pour représenter toute la complexité des langues humaines, mais, à l'inverse, plusieurs équipes montrent au cours des années 1990 que cette technologie est en fait extrêmement efficace, simple à mettre en œuvre et particulièrement appropriée pour la reconnaissance de séquences locales comme c'est le cas dans le cadre d'une analyse sémantique locale (voir notamment Hobbs et al., 1993).

Les années 2000 voient quant à elles se généraliser le recours aux systèmes fondés sur l'apprentissage. Ces systèmes sont en théorie plus portables que les précédents, car l'expert peut se contenter d'annoter un texte et c'est ensuite la machine qui « apprend une grammaire » ou, en tout cas, des règles permettant d'annoter les textes sur la base de l'annotation manuelle

Des résultats remarquables ont été obtenus ainsi, tant en qualité qu'en temps de développement (Tellier et Steedman, 2010 ; Gaussier et Yvon, 2011).

On voit aujourd'hui coexister les deux derniers types de systèmes. Le recours à l'apprentissage automatique reste un sujet de recherche et ce type de technique continue de se développer. Les entreprises commerciales ont quant à elles encore massivement recours aux systèmes à base de transducteurs à nombre fini d'états, notamment parce qu'ils offrent des qualités particulières (facilité de lecture et donc de révision par un humain ; les systèmes à base d'apprentissage artificiel sont beaucoup plus difficiles à corriger et à faire évoluer localement

Succès et difficultés

Comme on l'a vu, le traitement automatique des langues a permis des avancées majeures et est maintenant capable de fournir des modules efficaces pour traiter de grandes masses de données textuelles. Il faut toutefois souligner deux types de difficultés, ayant trait d'une part à l'analyse linguistique et de l'autre à l'ingénierie des connaissances.

En ce qui concerne l'ingénierie linguistique, on a souvent affaire à des architectures en « pipe-line » : un niveau d'analyse dépend du précédent (l'analyse sémantique repose sur l'analyse

syntaxique, qui elle-même repose sur la morphosyntaxe ou, pour prendre un exemple connexe, l'extraction d'événements dépend de l'analyse correcte de syntagmes exprimant des relations ou référant à des entités nommées). Chaque niveau d'analyse a tendance à amplifier les erreurs du niveau précédent, ce qui a évidemment une influence négative sur les performances globales. Par ailleurs, les systèmes sont peu performants dès que l'analyse dépasse les limites de la phrase, ce qui est pourtant souvent nécessaire, comme pour l'analyse des événements par exemple.

L'ingénierie des connaissances est une autre source de difficultés dès que l'on s'intéresse à des situations réelles. Par exemple, si on demande à un biologiste de reconnaître et d'annoter dans les textes des interactions entre gènes, celui-ci va avoir beaucoup de mal avec un grand nombre de cas parmi les plus importants. En effet, les cas avérés sont généralement exprimés clairement et sont relativement peu intéressants quand il s'agit d'informations connues, servant de « point de repère » (« le texte parle ici du gène X, qui a une interaction bien établie avec le gène Y »). Ce sont à l'inverse les cas limites qui sont précieux pour l'expert, mais problématiques pour l'analyse, car ils sont toujours exprimés avec des modalités et des prises de distance, ce qui ne permet pas de les catégoriser nettement. Un texte peut par exemple indiquer qu'il pourrait y avoir interaction entre deux gènes (clairement identifiés ou non) sans que l'auteur prenne position de façon certaine. Ce type de séquences est généralement très difficile à repérer pour le profane : on a souvent affaire à un discours très technique, modalisé, tout en nuances et qui emploie finalement peu des mots clés typiques identifiés pour la tâche. L'expert a aussi en général les plus grandes difficultés à annoter ces cas : face au linguiste, il va souvent avoir un discours complexe qui exprime une hésitation, qui « déroule » son raisonnement face au texte sans pouvoir toujours répondre de manière affirmative ou négative : « oui, il y a interaction » ou « non, ce n'est pas un cas d'interaction » (tout simplement parce que le texte ne prend pas position ainsi). Ce type de problème est aussi très présent en sciences sociales où les faits ne sont pas toujours « catégorisables » de façon claire : « annoter », c'est-à-dire « catégoriser », implique généralement de simplifier, ce qui peut s'opposer à la volonté de saisir un phénomène dans toute sa complexité.

Nous mettons ici volontairement le doigt sur les difficultés de ce type d'analyse. Il faut toutefois garder à l'esprit les grands succès obtenus depuis plusieurs années : les modules mentionnés supra montrent que des outils génériques, précis et performants existent pour plusieurs applications clés. La masse de données textuelles aujourd'hui disponible permet de concevoir des outils fondés sur l'apprentissage automatique, qui peuvent être adaptés très rapidement à un nouveau domaine si des données représentatives sont disponibles. L'utilisation de tels outils en sciences sociales est devenue quasi indispensable pour analyser l'information sur Internet ou sur les réseaux sociaux.