Traduisez ce texte 1 en Arabe

Analyse sémantique automatique

Cette section présente un aperçu du domaine et des techniques couramment employées aujourd'hui. Cette section comporte deux parties : la première consacrée à l'analyse factuelle (repérage d'entités nommées, de liens entre entités, etc.) tandis que la seconde partie sera consacrée à l'analyse dite subjective (analyse des sentiments, de l'opinion, etc.).

Analyse factuelle

Avant d'en venir à l'exposé des techniques actuelles, jetons un rapide coup d'œil à l'évolution du domaine, dans la mesure où l'historique permet de bien comprendre la situation présente.

Aperçu historique

Le traitement automatique des langues (TAL) est un domaine déjà ancien, qui est apparu dès les débuts de l'informatique, après la Seconde Guerre mondiale.

La première vague de développement du TAL (1945-1965) a mis en avant la traduction automatique. Cette application est extrêmement ambitieuse dans la mesure où, idéalement, elle suppose que la machine puisse « comprendre » le texte à traduire (suivant l'adage « pour pouvoir traduire un texte, il faut d'abord le comprendre ») avant de le « reproduire » dans la langue cible (on parle aussi de « génération » pour désigner cette deuxième phase). Cette ambition dépassait largement l'état de l'art dans les années 1950 et les grands espoirs initiaux d'avancées rapides n'ont pas donné les effets escomptés. Le rapport américain ALPAC (1966) a marqué le domaine, en donnant une vision très critique des expériences menées jusque-là, ce qui a abouti à un assèchement brutal des sources de financement, au moins du côté américain (Hutchins, 2001). Ce rapport critiquait essentiellement les approches trop naïves et surtout l'absence d'analyse approfondie des textes à traduire, ce qui laissait peu d'espoir de réel progrès dans le domaine à court ou moyen terme.

La période qui a suivi (1965-1985) s'est alors, logiquement, penchée sur la question de la compréhension automatique de texte. L'intelligence artificielle occupait une grande part dans ces travaux dans la mesure où la sémantique était mise en avant, ainsi que les formalismes de représentation des connaissances (pour prendre en compte notamment le lien entre connaissances linguistiques et connaissances sur le monde, cf. Sabah, 1988). Ces recherches, nombreuses et largement financées par des subsides publics aux États-Unis, avaient abouti, dans les années 1980, à un état de l'art peu lisible. Les recherches portaient sur des types de textes différents, avaient des objectifs divers et étaient rarement évaluées. Leur déploiement en milieu opérationnel semblait très lointain et les objectifs applicatifs encore flous et incertains, ce qui était évidemment un problème pour des agences de financement ayant avant tout des objectifs appliqués (Poibeau, 2003).

Les organismes de financement américains ont alors décidé de lancer des campagnes d'évaluation afin de rendre possible la comparaison de systèmes, en développant des tâches et des jeux de données publics, communs et réutilisables. Parallèlement, des métriques automatiques étaient mises au point afin de mesurer les performances, comparer les systèmes et leur évolution dans le temps. Les premières campagnes ont porté sur la compréhension de textes (conférences MUC, Message Understanding Conferences, 1987-1998) et ont été suivies par d'autres sur des thèmes similaires (TREC pour la recherche d'information, DUC puis TAC pour le résumé automatique, etc.).

Les premières campagnes, qui avaient un rôle exploratoire et laissaient une grande marge de manœuvre aux participants, ont montré que la compréhension de textes était en soi une tâche floue, complexe et mal définie. Qu'est-ce que comprendre un texte ? Comment formaliser cette notion ? Quel niveau de détail faut-il prendre en compte ? Les participants et les organisateurs se mettent alors d'accord, à la fin des années 1980, sur la nécessité de limiter dans un premier temps les ambitions au repérage d'informations factuelles, locales et faciles à évaluer. Les années 1990 verront ensuite l'apparition de sous-tâches particulières, menant au développement de modules de traitement génériques et réutilisables pour différents types d'applications.

Des modules d'analyse réutilisables

Les conférences MUC ont mis en avant un certain nombre de tâches (et/ou de modules d'analyse) qui sont fréquemment reprises pour les applications visant l'analyse de contenus en langage naturel (Poibeau, 2003, 2011).

Analyse des entités nommées : les entités nommées regroupent l'ensemble des séquences faisant référence à des entités connues, comme des personnes, des lieux, des entreprises ou des organisations. Par extension, les dates et les autres expressions numériques sont fréquemment regroupées avec les entités nommées. Les termes techniques sont aussi parfois assimilés à des entités, ce qui revient alors à élargir la classe à toutes les expressions d'intérêt pour un domaine donné.

Analyse de la coréférence : une même entité peut être dénommée de façon très variée dans un même texte (par ex. Jacques Chirac, le président Chirac, le président, il…). L'analyse de la coréférence vise à reconnaître les différentes dénominations d'une même entité, ce qui a un intérêt évident pour la compréhension de textes : on peut ainsi affecter à une même entité l'ensemble des informations qui la concernent, quelle que soit la forme sous laquelle cette entité apparaît en pratique.

Analyse des relations entre entités : cette tâche, au nom explicite, vise à identifier les relations entre entités telles qu'elles sont exprimées dans les textes. L'analyse de relations suppose une analyse correcte des prédicats, c'est-à-dire des éléments mettant en relation les différents éléments de la phrase (notamment les verbes et les noms prédicatifs) et, plus généralement, une analyse syntaxique correcte si on veut une analyse fiable et précise.

Analyse des événements : il n'y a pas de définition claire et précise de ce qu'est un événement, mais, au-delà de la simple analyse des relations, il est fréquemment nécessaire d'identifier des ensembles de plus haut niveau, rassemblant un certain nombre de relations simples dont l'agrégat est assimilé à un événement.

D'autres modules peuvent bien entendu être définis pour des besoins ou des tâches particulières. On a ainsi vu apparaître depuis quelques années une analyse plus fine des informations temporelles au sein des textes, ce qui est intéressant pour un grand nombre d'applications visant le suivi d'événements plus ou moins longs et leurs enchaînements. Les précédents modules semblent toutefois garder une plus grande généricité et être les plus communément repris au sein d'applications impliquant l'analyse de grandes masses textuelles.