Introduction

La révolution numérique de ces dernières décennies, née de la conjonction de la numérisation des données et de leur mise en réseau à un niveau planétaire, touche à l'un des fondamentaux de l'humanité, à savoir la communication, et affecte l'ensemble des activités humaines. Nous vivons aujourd'hui dans un monde numérisé et connecté, c'est une évidence, et nos modes de vie, qu'il s'agisse d'étudier, de travailler, de se divertir ou encore d'être citoyen, se sont en effet radicalement transformés. Cette société, dite de l'information, caractérisée à ses débuts par une augmentation sans précédent de la quantité de documents publiés, voit aujourd'hui s'abattre un véritable déluge avec encore plus de données, des contenus diversifiés et interactifs (avec le Web 2.0) et de nouveaux modes de publication et de partage des connaissances (avec le Web sémantique). Dans ce contexte, les systèmes de traitement de données permettent non seulement de les stocker, mais aussi de les manipuler. Un des objectifs poursuivis est alors d'extraire et de structurer des éléments d'information, à partir de données brutes, afin d'élaborer des connaissances et d'être en capacité de les exploiter. Le traitement automatique du langage (TAL) participe de cette entreprise pour ce qui est des données de nature langagière.

A cet égard, un des enjeux est notamment de capter l'information portée par les textes afin d'accéder à leur contenu. La tâche d'extraction d'information, formalisée dès la fin des années 1980, tente de répondre à ce besoin en s'attachant à reconnaître des éléments informationnels – quels qu'ils soient – dans les textes et à les mettre en relation les uns avec les autres. Parmi ces éléments figurent les *entités nommées* (EN), objets de cet ouvrage. En préambule, nous pouvons dire ici qu'il s'agit d'unités textuelles correspondant initialement à des noms de personnes, de lieux et d'organisations et dont le traitement s'articule en trois processus : *identification* ou recherche de ces unités dans les textes ; *catégorisation* ou typage selon des catégories sémantiques prédéfinies ; *liaison* ou processus de désambiguïsation permettant de résoudre la référence. Dès son apparition, la reconnaissance des entités nommées (REN) a connu un véritable succès, tant du point de vue des performances (en premier

lieu sur des textes anglais de nature journalistique) que des applications, devenant une brique élémentaire pour de nombreuses applications TAL.

Si les technologies de REN sont aujourd'hui relativement matures, les travaux et recherches sur les entités nommées poursuivent leur évolution. Il y a de nouvelles opportunités, avec notamment l'apparition de bases de connaissances volumineuses et multilingues, telle que Wikipédia, mais aussi de nombreux défis avec, entre autres : le traitement de ces unités pour les langues dites peu dotées, l'adaptation à de nouvelles formes d'écrits tels que présents dans les réseaux sociaux (Twitter), la reconnaissance selon des typologies plus complexes ou encore la désambiguïsation et la mise en relation (tâche d'*Entity Linking*). Aux côtés de ces nouvelles perspectives, il faut également définir plus avant de la notion d'entité nommée, améliorer les performances des systèmes pour les tâches existantes, affiner et faire évoluer les mesures d'évaluation adaptées à la tâche. Par ailleurs, le traitement des entités nommées semble susciter l'intérêt de nouvelles disciplines connexes, à l'instar des humanités digitales. Pour toutes ces raisons, le traitement des entités nommées demeure un domaine de recherche très actif, qui occupe une place centrale en TAL.

Le présent ouvrage s'intéresse à la tâche de REN pour le domaine général ; les domaines spécialisés (médical et biologique notamment) sont ainsi considérés comme hors du champ d'étude et ne seront que ponctuellement évoqués. Tout au long de cet ouvrage, nous considérons le texte écrit (possiblement issu de transcriptions) et utilisons des exemples en langue française et anglaise. Ces derniers sont extraits de textes journalistiques ou de guides d'annotations de programme de recherche (pour économie d'espace et lorsque cela n'est pas indispensable, ils sont parfois présentés sans contexte).

Ce volume vise à apporter les éléments utiles pour permettre au lecteur de se familiariser avec les notions centrales concernant les entités nommées et de découvrir les problématiques qui y sont liées, les méthodes disponibles pour les résoudre en pratique. En établissant un lien entre linguistique, informatique et besoins applicatifs, nous espérons que ce travail permettra d'apporter un éclairage sur toutes ces facettes simultanément, afin de mieux comprendre ce que sont les entités nommées et à quelles applications elles sont utiles. Pour ce faire, nous considérons, au travers de cet ouvrage, différents aspects concernant les entités nommées : comment est apparu le concept d'entités nommées, quelle est leur explication par la linguistique, quelles ressources permettent de les exploiter informatiquement, quelles méthodes sont utilisées pour les reconnaître et pour les lier à des référentiels, quelles sont les méthodes pour évaluer les systèmes qui les traitent.