Projet ANR blanc CONIQUE (2005-2008)

Inférences en contexte pour trouver, justifier et présenter des réponses à des questions en domaine ouvert.

Responsables : LIMSI (Brigitte Grau), CEA LIST (Olivier Ferret), MoDyCo (Jean-Luc Minel)

Dans le domaine de la recherche d'information, l'un des défis actuels porte sur la détermination de l'information précise cherchée par un utilisateur. L'objectif est de dépasser le paradigme de la recherche documentaire, dans lequel le système laisse à la charge de l'utilisateur le soin d'explorer une liste de documents pour y trouver l'information qu'il cherche, pour reporter la plus grande partie de ce travail sur le système de recherche d'information. Cette focalisation sur la recherche précise d'information s'est concrétisée ces dernières années par un intérêt porté aux systèmes de question-réponse en domaine ouvert. L'objectif de ces systèmes est de fournir une réponse à une question factuelle exprimée en langage naturel en trouvant cette réponse dans un ensemble de documents.

La plupart des systèmes extraient la réponse à une question en s'appuyant sur la sélection de passages de textes concentrant le plus grand nombre d'éléments de la question, y compris les éléments correspondant au type attendu de cette réponse. Ainsi, pour une question telle que

Quelle ville accueillit les Jeux Olympiques en 1992 ?

il est relativement aisé de sélectionner une phrase telle que

Les Jeux Olympiques d'été de 1992, Jeux de la XXVe olympiade de l'ère moderne, ont été célébrés à Barcelone …

et d'en extraire la réponse *Barcelone* en exploitant le fait que deux entités nommées de la question, *Jeux Olympiques* et *1992*, sont présentes dans la phrase et que celle-ci comporte une entité nommée de type ville correspondant au type de réponse attendu. Expliciter le lien entre *accueillit* et *ont été célébrés* n'est ainsi pas indispensable car les autres contraintes sont suffisamment fortes pour identifier la bonne réponse. Mettre en évidence de façon plus explicite les relations indirectes entre une question et une réponse, c'est-à-dire pouvoir justifier une réponse par rapport à une question, est en revanche nécessaire dans bon nombre de cas pour être capable de sélectionner les réponses correctes, comme l'illustre l'exemple suivant :

Quel chercheur allemand a reçu le prix scientifique Robert-Koch ?

Les prix seront remis le 23 octobre à l'Université de Bonn. Selon la Fondation Robert-Koch, le chercheur européen Peter Krammer et le scientifique japonais Shigekazu Nagata ont été distingués pour leurs recherches sur la mort génétiquement programmée des cellules apoptose.

- chercheur allemand ⇒ chercheur européen
- a reçu le prix ≡ ont été distingués
- fondation Robert-Koch ⇒ prix Rober-Koch

Le projet CONIQUE a pour objectif d'explorer cette problématique de la justification des réponses, qui fait par ailleurs l'objet d'une piste spécifique de CLEF-QA, Answer Validation Exercice (AVE), depuis 2006. Cette exploration se fait selon une perspective répondant aux exigences d'un travail en domaine ouvert en se focalisant plus particulièrement sur l'inévitable incomplétude des bases de connaissances utilisées. Le projet n'a donc pas pour but de constituer ou d'exploiter une base de connaissances a priori permettant de répondre aux questions et de justifier les réponses trouvées. Il se concentre au contraire sur la modélisation de l'extraction de ces connaissances à partir de différents textes en fonction des besoins nécessaires à la construction d'un chemin inférentiel entre les éléments trouvés dans les textes et l'information cherchée, telle qu'elle

est spécifiée par une question. En outre, l'extraction des connaissances à partir de textes présente l'intérêt de pouvoir disposer en parallèle du contexte d'usage et de validité de ces connaissances. Ce contexte est particulièrement important pour contrôler l'enchaînement des inférences et leur validité dans le processus de recherche d'une réponse mais il est aussi très intéressant pour la présentation des réponses. Présenter les réponses possibles à une question en les accompagnant de leur contexte (date, lieu, point de vue…) permet en effet à l'utilisateur de comprendre l'origine de leurs différences.

Tâches réalisées :

- *Justification du contexte temporel :* Un repérage automatique des expressions calendaires faisant référence à un grain usuel, soit de manière explicite (ex : le 3 janvier 2002, en mars 1987, depuis avril, etc.), soit de manière implicite (ex : en 2007), a été effectué. Une comparaison entre expression calendaire de la question et du passage candidat permet de dire si la référence temporelle de la réponse proposée est identique, inclus, plus large ou différente de celle de la question.

  D. Battistelli, J.-L. Minel, S. R. Schwer : Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. TAL vol 47, n° 3, 2007.

  D. Battistelli, J. Couto, J.-L. Minel, S. R. Schwer : Représentation algébrique des expressions calendaires et vue calendaire d'un texte, Actes *TALN'08* (*Traitement automatique du langage naturel 2008*), 8-12 juin 2008, Avignon.

- *Système de décision :* La décision de dire si une réponse est justifiée ou non est prise par apprentissage à partir des caractéristiques de la question qui sont reconnues dans le passage.

  A.-L. Ligozat, B. Grau, A. Vilnat, I. Robba, A. Grappy, Lexical validation of answers in Question Answering, 2007 IEEE / WIC / ACM international conference on Web Intelligence (WI 07), 2007

- *Justification des relations entre éléments de la question*
  L'identification des relations est réalisée par analyse syntaxique. Le repérage et l'identification automatique de relations sont étudiés. Un autre travail en cours produit un ensemble de relations instanciées pour représenter les questions. Ces informations sont recherchées par réécriture dans les passages candidats et les informations manquantes (donc non encore justifiées) sont recherchées dans d'autres documents, par réinterrogation.

  M. Embarek et O. Ferret. Une expérience d'extraction de relations sémantiques à partir de textes dans le domaine médical. 14ème conférence TALN 2007, p. 37-46, 2007.

- *Constitution d'un corpus d'évaluation*
  Il contient des questions, les réponses avec le document dont elles sont extraites, en indiquant le type de raisonnement nécessaire, à savoir résolution de référence, vérification du type de la réponse, chaîne d'inférence sémantique, contexte temporel ou spatial.