Recherche d'information précise dans des sources d'information structurées et non structurées : défis, approches et hybridation

Brigitte Grau* — **Anne-Laure Ligozat*** — **Martin Gleize****

* *LIMSI, CNRS, ENSIIE, Université Paris-Saclay, Bât 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay*
** *LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Bât 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay*
prenom.nom@limsi.fr

RÉSUMÉ. *Cet article propose une synthèse d'une part sur les approches développées en questions-réponses (QR) sur du texte, en insistant plus particulièrement sur les modèles exploitant des représentations structurées des textes, et d'autre part sur les approches récentes en QR sur des bases de connaissances. Notre objectif est de montrer les problématiques communes et le rapprochement possible de ces deux types de recherche de réponses en prenant appui sur la reconnaissance des relations présentes dans les énoncés textuels et dans les bases de connaissances. Nous présentons les quelques travaux relevant de ce type d'approche afin de mettre en perspective les questions ouvertes pour aller vers des systèmes réellement hybrides ancrés sur des représentations sémantiques.*

ABSTRACT. *This paper provides a synthesis of text-based Question Answering (QA) approaches, with emphasis on models exploiting structured representations of texts, and recent QA approaches for knowledge bases. Our goal is to show the common issues, and to identify what can be a common approach, using both the relations from textual documents and the triplets from knowledge bases. We present a few works using this type of approach in order to highlight open issues and pave the way to really hybrid systems based on semantic representations.*

MOTS-CLÉS : *questions-réponses sur des textes, questions-réponses sur bases de connaissances.*

KEYWORDS: *Textual Question-Answering system, Knowledge-Base Question Answering system.*

TAL. Volume 56 – n°3/2015, pages 75 à 99

1. Introduction

La recherche d'information couvre un vaste éventail de tâches, dans le but de répondre à des besoins utilisateurs différents. Lorsqu'il s'agit de rechercher des documents, ce besoin est généralement exprimé par une liste de mots-clés, tandis que pour rechercher une information précise, concernant un fait ou une entité, ce besoin peut être formulé simplement par une question en langage naturelle, comme par exemple « Dans quelle ville est né l'assassin de Martin Luther King ? ». Le fait de partir d'une telle requête permet d'exploiter tous les outils d'analyse de la langue, et d'en réaliser une analyse syntaxique et sémantique assez poussée.

La nature des informations recherchées conduit aussi à développer des processus de recherche différents. Selon que l'on interroge des textes, qui sont des sources de connaissances non structurées, ou des bases de connaissances, sources de connaissances structurées, les approches envisagées peuvent être très différentes. La recherche de réponses dans des textes est fondée sur un appariement ou une similarité entre la question et les extraits de texte sélectionnés constituant des passages réponses. Les informations extraites de la question permettent de formuler des contraintes pour cet appariement ou d'établir des critères de similarité. La recherche de réponses dans une base de connaissances suppose la reconnaissance et l'identification des entités et des relations qui les lient, pour construire une requête en un langage formel et interroger la base. L'ensemble des relations, celui du schéma de la base, est alors connu.

À une extrémité du spectre, en questions-réponses (QR) sur du texte, les processus s'appuient sur des informations de surface qui sont combinées par des méthodes statistiques ; à l'autre, en QR sur une base de connaissances, les processus exploitent, comparent ou transforment des représentations sémantiques structurées sous forme de graphes d'entités et de relations pour produire une requête dans un langage formel. On voit ainsi apparaître deux domaines de recherche à chaque extrémité, qui sont la recherche d'information et l'interrogation de bases de connaissances du Web sémantique, les deux s'appuyant sur une analyse de la langue. Néanmoins, le problème peut être posé sous un même point de vue, conduisant à une formalisation plus unifiée : les textes peuvent être vus comme des bases de connaissances, dans la mesure où ils contiennent des relations entre éléments d'information, qui peuvent être obtenues par analyse syntaxique et sémantique. Cela est d'autant plus justifié que les systèmes de QR sur du texte qui s'appuient sur une représentation structurée des phrases sont souvent plus performants.

Le bénéfice attendu d'une modélisation unifiée de ces deux tâches est de pouvoir les exploiter conjointement pour rechercher une information précise et étendre l'éventail des réponses possibles, puisqu'elles pourraient être issues de textes ou de bases de connaissances. Ainsi, pour la question portant sur la ville de naissance de l'assassin de Martin Luther King, la réponse peut être difficile à extraire d'un seul passage, qui a peu de chances de contenir toutes les informations de la question, ou d'une base de connaissances de type DBpedia, qui ne contient pas la relation entre Martin Luther King et son assassin. En revanche, « James Earl Ray, l'assassin de Martin Luther

King », peut être extrait de textes, indépendamment de sa ville de naissance, et sa ville de naissance peut résulter de l'interrogation de DBpedia.

Le présent article propose une synthèse d'une part sur les approches développées en QR sur du texte, en insistant plus particulièrement sur les modèles exploitant des représentations structurées des textes (section 3) et, d'autre part, sur les approches récentes en QR sur des bases de connaissances (section 4), après avoir présenté les problématiques traitées et le rapprochement possible de ces deux types de recherche de réponses (section 2). Nous discuterons les problèmes et possibilités ouverts par la collaboration d'informations issues des deux types de ressources (section 5). Nous présenterons les quelques travaux relevant de ce type d'approche afin de mettre en perspective les questions ouvertes pour aller vers des systèmes de questions-réponses réellement hybrides ancrés sur une représentation sémantique des informations.

2. Problématique globale

La recherche d'information dans des bases de connaissances ou des bases de données, à partir de questions posées en langage naturel est ancienne. Dès les années 60, de tels systèmes ont été développés (voir (Barr et Feigenbaum, 1981) pour une description de ces systèmes). Il en est de même de la recherche de réponses à des questions dans des textes (Lehnert, 1977). La principale limitation de ces systèmes résidait dans leur application à un domaine restreint, pouvant être circonscrit et représenté formellement, ce qui a conduit à l'abandon de ces recherches. Ce n'est que récemment que l'on a vu renaître un intérêt pour ces tâches avec la disponibilité d'outils d'analyse de la langue largement applicables, de ressources telles que WordNet (Miller, 1995) ou des bases de paraphrases (Ganitkevitch *et al.*, 2013), ainsi que l'essor du Web qui a conduit à la constitution de bases de connaissances en domaine ouvert.

Nous nous intéressons dans cet article à la recherche de réponses en domaine ouvert. En effet, la recherche de réponses en domaine de spécialité, sur le texte et sur des bases de connaissances, repose sur des méthodes qui exploitent des sources de connaissances structurées, de type ontologie ou taxonomie de domaine, et rejoignent les méthodes sur le texte qui nous intéressent ici.

2.1. *Évaluations et corpus*

En 1998, est apparue à TREC [1] la tâche de questions-réponses. En 2002, la tâche a été introduite à NTCIR [2] et en 2003 à CLEF [3], attestant ainsi de son intérêt. Elle a été reconduite jusqu'en 2007 à TREC et 2009 à CLEF. L'évaluation consiste, étant donné un ensemble de questions, à évaluer la réponse donnée par les systèmes à

1. *Text REtrieval Conference*, `http://trec.nist.gov/`
2. *NII Testbeds and Community for Information access Research*, `http://ntcir.nii.ac.jp/`
3. *Conference and Labs of the Evaluation Forum*, `http://www.clef-initiative.eu/`

chacune des questions, celle-ci étant constituée de la réponse courte associée à un document ou un passage permettant de la justifier. La mesure d'évaluation est alors le nombre de réponses correctes. Lorsque les systèmes peuvent retourner plusieurs réponses, entre 3 et 5 généralement, la mesure d'évaluation est le MRR (*Mean Reciprocal Rank*). L'évaluation globale de systèmes de questions-réponses sur le texte a laissé la place à l'évaluation de processus plus ciblés, ce qui permet de participer à ces tâches sans avoir à développer un système complet, et de pouvoir se focaliser sur certains processus. La campagne RTE[4], qui a eu lieu de 2006 à 2011, a permis de cibler la tâche d'implication textuelle (Dagan *et al.*, 2006) ; il en est de même pour la tâche de CLEF QA4MRE@CLEF [5], existant depuis 2011 et qui consiste à répondre à des QCM sur des textes, tâche pouvant se ramener à de l'implication textuelle. Les campagnes d'évaluations SemEval proposent depuis récemment des tâches d'implication textuelle pour l'évaluation automatique de réponses d'étudiants (Dzikovska *et al.*, 2013), et une tâche appelée similarité sémantique textuelle (STS) dont le but est de noter de 0 à 5 l'équivalence sémantique de deux énoncés courts (Agirre *et al.*, 2012).

Grâce au travail mené autour de Wikipédia et la création de DBpedia (Lehmann *et al.*, 2014), l'évaluation de systèmes d'interrogation d'une base de connaissances en langage naturel est proposée dans le cadre de QALD@CLEF [6] depuis 2011. QALD propose différentes tâches, dont une requérant le recours à des processus hybrides. Les requêtes sont posées sur DBpedia, et les réponses peuvent être uniques ou formées d'une liste. Chaque jeu de test comporte 60 questions. Dans le même courant, la création de FreeBase (Bollacker *et al.*, 2008) par les utilisateurs du Web, dont le contenu est maintenant transféré dans Wikidata [7], a favorisé la proposition de différentes approches afin de faciliter son interrogation. En domaine médical, BioAsq [8] comporte une tâche de questions-réponses, sur le texte mais aussi sur des ontologies.

L'existence de ces travaux a conduit à la création de jeux de données. En plus des ressources des campagnes citées, plusieurs jeux de questions ont été mis à disposition par des équipes de recherche : Cai et Yates (2013) ont créé un corpus de 917 questions associées à leur représentation en lambda calcul utilisant les relations Freebase ; Berant *et al.* (2013) ont créé le corpus WEBQUESTIONS d'environ 6 000 questions à partir de l'API de Google suggest, comprenant également les entités Freebase répondant à chaque question.

2.2. *Tâche de questions-réponses*

Les systèmes de QR suivent généralement l'architecture présentée figure 1.

4. *Recognizing Textual Entailment*, tâche de la campagne TAC (*Text Analysis Conference*), `http://www.nist.gov/tac/`
5. *Question Answering for Machine REading*, `http://nlp.uned.es/clef-qa`
6. `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/`
7. `https://www.wikidata.org/wiki/Wikidata:Main_Page`
8. `http://www.bioasq.org/participate/challenges`

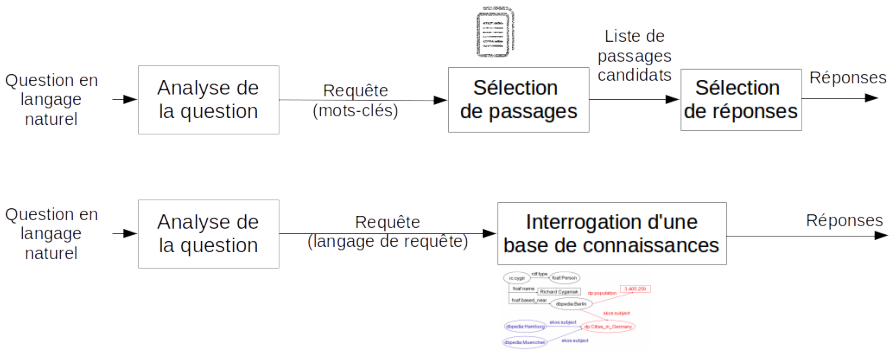

Figure 1. *Architecture de recherche de réponses à partir de questions en langage naturel*

Quelle que soit la ressource interrogée, la première étape consiste à analyser la question en langage naturel afin d'en tirer toutes les informations exploitées par les processus de recherche de passages et de sélection de réponses.

La recherche de réponses dans des textes procède ensuite à une sélection de passages dans les documents interrogés afin de restreindre l'espace de recherche, en mettant en œuvre des méthodes de recherche d'information. Un passage de texte est pertinent s'il contient l'information donnée dans la question et la réponse courte attendue. Généralement, cette information n'est pas exprimée dans les mêmes termes que ceux de la question et différents types de variations linguistiques sont à traiter entre la question et les passages. Le problème de sélection d'une réponse peut être modélisé par une implication textuelle de la forme $T => H$, dont la signification est la suivante (Glickman, 2006) : un lecteur humain lisant T peut raisonnablement en déduire H. Dans le cas de la recherche de réponses, le texte T est un passage sélectionné, et l'hypothèse H correspond à une formulation déclarative de la question dans laquelle la place de la réponse attendue est marquée et éventuellement instanciée par la réponse que l'on cherche à valider. C'est cette étape qui nous intéresse plus spécifiquement et pour laquelle nous présenterons en section 3 les méthodes proposées.

La recherche de réponses dans une base de connaissances nécessite la construction de requêtes dans un langage formel représentant la question. Celle-ci peut être représentée sous forme d'un graphe connexe de relations de la base partiellement instanciées, dont l'une des variables à instancier est la réponse. Le problème des variations lexicales entre labels associés aux entités et relations de la base et les termes employés par l'utilisateur se pose alors, puisque ce dernier n'est pas guidé par la connaissance du schéma de la base. Se pose également le problème de la résolution d'ambiguïtés sémantiques, car un même terme peut faire référence à différents éléments sémantiques.

Illustrons ces processus avec la question suivante : « Who is the daughter of Bill Clinton married to ? » La réponse peut être trouvée dans les passages suivants :

> On 31 July 2010, 30-year-old *Chelsea Clinton* (the daughter of former U.S. president Bill Clinton and Secretary of State Hillary Clinton), who had recently received a master's degree from Columbia University's Joseph L. Mailman School of Public Health, married 32-year-old *Marc Mezvinsky*, an investment banker.

Dans ce premier passage, les mots de la question se retrouvent à l'identique, mais en revanche ils sont séparés par deux incises, ce qui rend la structure syntaxique nécessaire pour les relier.

> Famously protective of her private life, the 32-year-old daughter of Bill and Hillary Clinton has long stayed silent on her relationship with husband *Marc Mezvinsky*.

Dans ce second passage, la relation de mariage est exprimée cette fois par une variation sémantique, le nom « husband », à relier au mot « married » de la question, et le nom de Bill Clinton est associé à celui de sa femme, ce qui rend son identification plus complexe. En outre, cet exemple montre qu'il n'est pas nécessaire d'identifier la fille de Bill Clinton pour trouver la réponse.

La réponse peut aussi être trouvée dans DBpedia par la requête indiquée dans la figure 2. Afin de former cette requête, il faut repérer l'existence d'une entité implicite correspondant à Chelsea Clinton (?a dans la requête), relier l'entité Bill Clinton à la ressource correspondante, et identifier les relations « spouse », correspondant à « married », et « child » correspondant à « daughter ». Il faut également pouvoir relier ces différentes ressources entre elles pour former les relations, puis la requête complète.

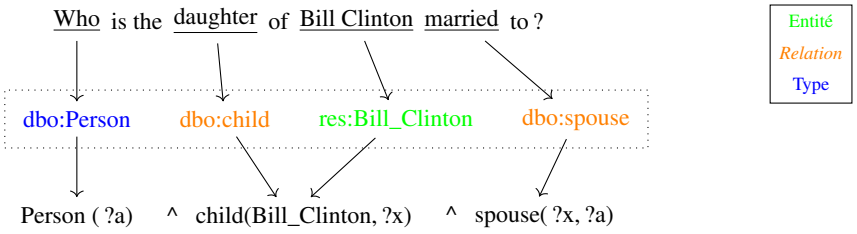

Figure 2. *Analyse de la question et construction de la requête*

2.3. *Problématique traitée*

Sur les textes, la recherche de réponses à des questions pose le problème de comparer le sens de la question avec le sens d'un passage réponse. Sur une base de connaissances, la recherche de réponses pose le problème du passage d'un texte à un langage formel. Dans les deux paradigmes, les représentations manipulées peuvent être sous-tendues par la notion de relation. C'est pourquoi nous allons poser les définitions nécessaires à une mise en parallèle des problématiques.

2.3.1. *Définitions*

Une *entité* e_1 représente un objet du monde, avec un identifiant unique, par exemple son *uri* dans une base de connaissances. Elle est liée à un type, ou une classe, et est référée dans les textes par différentes séquences de mots, correspondant à des *mentions*. Les entités nommées reconnues dans un texte, par exemple des personnes ou des lieux, sont des mentions d'entités auxquelles on a associé un type.

Une *relation* R associe des entités d'un domaine $D1$ à des entités d'un domaine $D2$. Un domaine correspond à une classe, ou à un ensemble de classes. Une *instance* de relation associe une entité $e1$ à une entité $e2$ par une relation R. Une *mention* de relation est la réalisation phrastique d'une instance de relation dans un énoncé, permettant de lier des mentions d'entités.

Par exemple *auteur_de(Personne, Œuvre)* est une relation, *auteur_de(Daniel_Defoe, Robinson_Crusoé)* est une instance de la relation *auteur* et « Daniel Defoe a écrit Robinson Crusoé » est une mention de la relation, et par extension « a écrit ».

Une *base de connaissances* est constituée d'entités reliées entre elles par des relations binaires ou n-aires. Ces bases de connaissances ne sont pas toutes structurées de la même façon : DBpedia, YAGO et MusicBrainz contiennent des triplets (donc des relations binaires) alors que Freebase contient des n-uplets (et donc des relations n-aires). Il est néanmoins possible de passer d'une représentation à l'autre. Le langage standard d'interrogation de ces bases de connaissances est SPARQL, langage proche du SQL, qui permet d'interroger les données RDF des ressources du Web sémantique.

2.3.2. *Mise en parallèle des deux tâches*

Les méthodes permettant de résoudre les deux tâches sont diverses et pour l'heure très différentes. Néanmoins, certaines ont en commun une représentation des informations qui prend appui sur la notion de relation, pour la question, la base de connaissances et les passages. Nous présentons ici une mise en parallèle des deux tâches, sur laquelle nous reviendrons après avoir exposé les méthodes existantes. Il est important de constater que la mise en parallèle de ces deux tâches ouvre la voie vers la possibilité de définir un formalisme de représentation des informations qui soit applicable dans les deux tâches et qui permette la recherche hybride de réponses.

Nous représentons dans la figure 3 la question comme un ensemble de mentions de relation ($m_i r_j$) entre mentions d'entités ($m_i e_j$). La réponse recherchée est une entité particulière indiquée par ? et certaines relations ou entités peuvent ne pas être explicites (x). L'interrogation de textes pose le problème de comparer une représentation comportant des mentions de relation et entités de la question avec différentes mentions de ces mêmes entités et relations, présentes dans les passages pertinents. Puisque l'on ne cherche pas à identifier des relations sémantiques mais à reconnaître des paraphrases, la définition de ce qui est mention de relation et entité n'est pas assujettie à une représentation des connaissances explicite. De plus, la même information est souvent donnée sous différentes formes, ce qui ne rend pas toujours nécessaire de

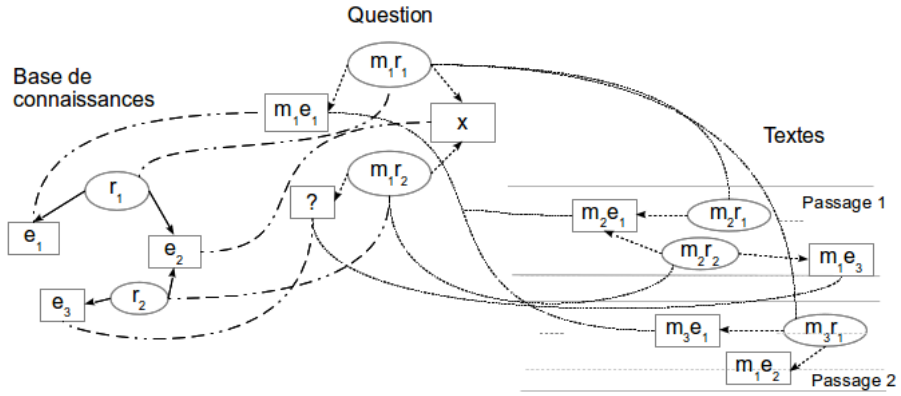

Figure 3. *Recherche de réponses dans une base de connaissances ou dans les textes*

reconnaître toutes les variantes d'une même information pour trouver une réponse car généralement on ne recherche pas toutes ses occurrences.

L'interrogation d'une base de connaissances (cf. figure 3) pose le problème de la transformation d'une structure comportant des mentions de relation et d'entité de la question vers une structure, *e.g.* un graphe, formée de l'instanciation partielle des relations de la base. Cela impose de savoir reconnaître toutes les variations présentes dans la question. Cependant, la représentation formelle des relations fournit des contraintes sémantiques pour reconnaître les arguments d'une relation, et permet de concevoir un processus d'identification des instances de relation plus informé.

Nous présentons dans les deux sections suivantes les méthodes utilisées pour répondre à des questions, d'abord sur les textes puis sur les bases de connaissances.

3. Recherche de réponses dans des textes

Ainsi que nous l'avons énoncé précédemment, la sélection de réponses se ramène à un problème d'implication textuelle. Il ne s'agit pas seulement de trouver que deux formulations sont analogues, ce qui correspondrait à une stricte paraphrase, mais de pouvoir différencier les informations figurant dans le passage qui ne gênent pas l'implication textuelle de celles qui empêchent cette implication et amènent à des contradictions. Nous verrons que la plupart des modèles n'intègrent pas explicitement la reconnaissance de contradictions.

La sélection de réponses a donné lieu à la proposition de nombreuses méthodes qui s'appuient sur des niveaux d'analyse de la langue différents, et donc des représentations différentes des énoncés (questions et passages) :

– représentation par « sac de mots ». Les énoncés sont caractérisés par différents traits liés au lexique ;

– représentation de séquences de termes. Aux informations précédentes, sont ajoutées des informations sur l'ordre des mots ;

– représentation lexicale structurée : arbre syntaxique ou syntaxico-sémantique, tenant compte de la variabilité lexicale ;

– représentation conceptuelle : prédicats, graphes de concepts.

Quelles que soient les représentations choisies, elles ont donné lieu à la proposition de méthodes nécessitant plus ou moins de supervision. Par ailleurs, la plupart des méthodes récentes traitent la sélection de passages pertinents, mais pas l'extraction de la réponse courte. De ce fait elles ne sont généralement pas évaluées sur les mêmes jeux de données que les systèmes de QR complets. Les travaux les plus récents évaluent leurs propositions sur le jeu de phrases constitué par Wang *et al.* (2007), présenté section 3.2, qui évalue en fait le classement ou la sélection de phrases réponses et non pas le processus de questions-réponses, malgré ce que la terminologie dans une part de la littérature sur le sujet peut laisser penser.

3.1. *Représentations de type « sac de mots » et séquences de termes*

De nombreuses approches ont cherché à définir une mesure de similarité entre un passage réponse et la question (Ferret *et al.*, 2001 ; Magnini *et al.*, 2002 ; Ittycheriah *et al.*, 2001) et certaines reposent sur un apprentissage supervisé, formulant la sélection de passages comme un problème de classification ou de réordonnancement (Suzuki *et al.*, 2002 ; Grappy *et al.*, 2011). Aux informations lexicales, telles que mots communs, variations sémantiques, types d'entités nommées, peuvent être ajoutés des traits syntaxiques, tels que les relations de dépendances syntaxiques communes (Volokh et Neumann, 2010 ; Moriceau *et al.*, 2009). La cohésion de l'ensemble est mesurée par des notions de densité ou des calculs de la plus grande chaîne commune (Kozareva *et al.*, 2006 ; Grappy *et al.*, 2011).

Afin d'introduire l'ordre des mots dans le calcul de similarité, on peut représenter la question et le passage par des séquences de mots ou de syntagmes et calculer des distances d'édition sur des *chunks*, *i.e.* des termes multimots, (Bernard *et al.*, 2010), ou rechercher un alignement des deux séquences. Il est possible ainsi de capturer des paraphrases syntaxiques, sans annotation préalable. Bu *et al.* (2012) et Gleize et Grau (2015) utilisent ainsi des fonctions noyau mesurant la similarité de deux paires de séquences de mots en terme de règles de réécriture communes dans un SVM. Ces systèmes rivalisent avec l'état de l'art pour la reconnaissance d'implication textuelle.

Ces approches montrent des limitations lorsqu'il est nécessaire d'éviter des interprétations erronées qui peuvent être induites en se fondant sur ces représentations, notamment dans le cas de présence d'incises ou de relations de dépendances distantes. Par exemple, à la question « Qui a assassiné Henri IV ? », ne pas sélectionner le passage « Henri III, comme Henri IV, avait été assassiné par un fanatique, Jacques Clé-

ment. », requiert de raisonner sur une représentation complète de l'énoncé permettant d'identifier qui a tué qui.

3.2. *Représentations lexicales structurées*

Afin d'étudier plus spécifiquement le rôle de la syntaxe pour sélectionner des réponses, plusieurs travaux ont proposé des méthodes de comparaison de représentations issues d'une analyse syntaxique pour sélectionner des passages pertinents. Ces travaux ont pris majoritairement naissance lors des campagnes RTE. L'analyse en dépendances est généralement préférée à celle en constituants, pour la structure de graphe obtenue. La disponibilité d'un bon analyseur syntaxique est requise et est cependant plus limitée selon les langues : seules les langues bien dotées disposent de très bons outils (Klein et Manning, 2003 ; Nivre *et al.*, 2007 ; Chen et Manning, 2014).

L'appariement de deux représentations syntaxiques doit prendre en compte les différences de formulation syntaxique d'un même contenu pour aligner les relations entre mots ou syntagmes analogues. Il s'agit donc d'apparier à la fois des items et les relations qu'ils entretiennent. Les modèles qui obtiennent les meilleurs résultats intègrent des connaissances syntaxiques et sémantiques. De ce fait, la représentation des énoncés n'est pas purement syntaxique, mais fondée sur leur représentation syntaxique, augmentée par des connaissances sémantiques lexicales plus ou moins intégrées au modèle.

Cui *et al.* (2005) cherchent à aligner chaque relation de dépendance présente dans la question avec un chemin de dépendances dans la phrase candidate, et leur associent un score correspondant à la probabilité de transformer une relation en un chemin, fondé sur l'information mutuelle. Intégré à un système de QR, cette approche produit un apport significatif sur les questions de TREC12 (P@1= 0,38) par rapport à un appariement strict des relations de dépendance, et une approche de base de type « sac de mots ». Ce type d'approche a été amélioré sur la tâche de sélection de phrases réponses par des modèles qui cherchent à apprendre l'alignement latent entre les deux représentations. Différentes mesures ont été conçues reposant sur un décompte des alignements entre les relations de dépendance des deux fragments de texte. Wang *et al.* (2007) apprennent cet alignement par un modèle probabiliste génératif fondé sur une grammaire quasi synchrone. Ce formalisme permet de tenir compte de transformations locales, entre des équivalences de syntagmes tels que « leader of France » et « French president ». Ces modèles réalisent un alignement mot à mot où un mot de la question est associé à un mot du passage, ce qui ne permet pas de trouver des paraphrases sous-phrastiques de tailles différentes ; ceci a conduit Yao *et al.* (2013) à étendre un CRF par un modèle semi-markovien qui permet d'introduire cette flexibilité.

L'édition d'arbres de dépendances pour une tâche de sélection de passages a été proposée pour la première fois par Punyakanok *et al.* (2004). Elle a été intégrée à un système de validation de réponses par Kouylekov *et al.* (2007). Heilman et Smith (2010) en ont proposé une version évaluée sur la sélection de phrases réponses, qui

apprend un classifieur sur des traits extraits des séquences d'opérations d'édition. Celles-ci sont produites par un algorithme glouton appliquant une fonction noyau sur les arbres pour évaluer leur similarité. Dans (Moschitti *et al.*, 2007), les auteurs proposent une fonction noyau sur les arbres qui étend la définition de sous-arbres similaires. Enfin, le modèle de Wang et Manning (2010) apprend l'alignement latent de deux arbres de dépendances en définissant des opérations d'édition qui incluent des transformations sémantiques, intégrant ainsi syntaxe et sémantique.

Ces modèles sont évalués sur la sélection de phrases [9]. Le jeu de test de (Wang *et al.*, 2007) contient 89 questions et 1 517 phrases extraites des données de TREC13. Les phrases sont annotées manuellement positivement si elles contiennent la réponse. Cet ensemble, avec une moyenne par question de 17 phrases candidates filtrées pour contenir moins de 40 mots chacune, ne représente pas la même diversité que ce que l'on trouve dans les jeux de test de la tâche questions-réponses à TREC.

Une limitation de ces modèles porte sur le fait que les représentations à aligner sont des phrases, alors que souvent les informations de la question sont réparties sur plusieurs phrases. Dans (Sachan *et al.*, 2015), les auteurs proposent un modèle d'alignement intégrant, entre autres, cette contrainte. Il a été conçu pour la tâche de compréhension de textes, *Machine Reading*, et est évalué sur le corpus MCTest[10] qui est un ensemble de QCM destinés à évaluer la compréhension d'un texte, ce qui ne permet pas d'évaluer son apport dans une tâche de questions-réponses.

Les relations syntaxiques rendent compte des relations sémantiques entre termes, mais il faut gérer les différentes formes de relations syntaxiques possibles en plus de réaliser les inférences sémantiques pour établir une implication textuelle. Afin de séparer les deux problèmes et modéliser uniquement le processus d'inférence, quelques modèles s'appuient sur une représentation conceptuelle.

3.3. *Représentations conceptuelles*

Plusieurs travaux ont cherché à représenter le texte et l'hypothèse en expression logique obtenue à partir d'analyses syntaxiques et sémantiques profondes afin de prouver que le texte implique ou non l'hypothèse. Ce processus met en général en œuvre une notion de relâchement lors de la construction de la preuve afin d'être robuste.

Le système COGEX (Tatu *et al.*, 2006) cherche à justifier des réponses en recherchant une chaîne d'inférences entre hypothèse et texte. L'appariement entre une question et une phrase consiste à prouver une formule logique constituée des axiomes provenant de la phrase, de la question niée pour effectuer une preuve par l'absurde, et d'axiomes permettant de relier des informations entre elles (entités nommées, types de réponses, possessifs, etc.). Par ailleurs, le démonstrateur utilise si nécessaire des

9. Voir `http://aclweb.org/aclwiki/index.php?title=Question_Answering_State_of_the_art` pour obtenir les scores des différents systèmes.
10. http ://research.microsoft.com/mct

chaînes d'inférences reliant des synsets de WordNet, provenant de la ressource Extended WordNet (Mihalcea et Moldovan, 2001). Si la réfutation n'est pas détectée, une étape de relaxation est appliquée en retirant des prédicats. Chaque prédicat retiré ayant un poids, il est possible de calculer un poids global qui sera comparé à une valeur seuil afin de détecter l'implication potentielle. C'est donc une méthode hybride qui combine un mécanisme par preuve et l'importance des différents termes de la question. Le système de QR intégrant ce modèle avec des modèles plus robustes, a obtenu les meilleurs résultats aux campagnes TREC (Moldovan *et al.*, 2003).

Glöckner *et al.* (2007) mettent aussi en œuvre un mécanisme par preuve logique mais débutent par une étape visant à normaliser le texte en modifiant les mots afin qu'ils soient au plus près de ceux de l'hypothèse. La preuve est effectuée par un mécanisme de relaxation récursive en retirant des prédicats de l'hypothèse à l'ensemble des prédicats de départ issus du texte et de l'hypothèse. La décision finale est prise à partir des mots présents dans cette nouvelle hypothèse. Le système de QR intégrant ce processus a obtenu d'excellents résultats à CLEF en 2005 et 2006 (Hartrumpf, 2006). Ces approches font usage de représentations sémantiques conceptuelles, c'est-à-dire des ressources représentant des concepts, leur définition et leurs relations, qui sont construites manuellement et qui sont rares.

En revanche, la production d'analyses sémantiques de surface est désormais possible grâce aux récents progrès effectués en identification des rôles sémantiques. Il est ainsi permis de s'abstraire des relations syntaxiques et de leurs variations pour exploiter des relations entre constituants. Ces progrès sont essentiellement dus à l'existence de ressources comme FrameNet [11] et PropBank [12]. Dans (Shen et Lapata, 2007), les structures sémantiques de la question et de la phrase candidate construites à partir de FrameNet sont comparées par un modèle d'appariement de graphes fondé sur la comparaison des sous-graphes constitués des prédicats instanciés de FrameNet, et la phrase de meilleur score est conservée. Ce module améliore les résultats d'un système de QR, lorsque les ressources liées à la question existent, ainsi qu'associé à une stratégie fondée sur les relations syntaxiques en l'absence de couverture de la ressource. Sammons *et al.* (2009) utilisent l'étiqueteur de rôles sémantiques de l'Illinois (Punyakanok *et al.*, 2008) pour annoter des couples texte et hypothèse du problème de RTE. Les structures de graphes prédicat-arguments de PropBank obtenues sont moins morcelées et profondes que des dépendances syntaxiques, mais sont aussi moins précises sémantiquement que les cadres de FrameNet. Leur système produit d'abord des scores d'appariement à l'aide de métriques, qui peuvent être définies sur toutes les couches d'annotation qu'ils utilisent, parmi lesquelles les segments, les entités nommées et les structures de PropBank. Ensuite, plusieurs alignements pour chacune des couches sont produits par optimisation sous contraintes. Enfin, des traits sont extraits de ceux-ci et utilisés dans un classifieur SVM. Un algorithme est proposé pour apprendre les alignements en utilisant le classifieur dans une boucle de rétroaction. Inclure les traits de structure sémantique améliore les résultats sur le corpus RTE5.

11. https ://framenet.icsi.berkeley.edu
12. https ://verbs.colorado.edu/propbank/

3.4. *Conclusion sur la recherche de réponses dans des textes*

3.4.1. *Problèmes étudiés*

À l'heure actuelle, peu de systèmes de QR sur le texte sont encore développés et publiés. Pour les plus aboutis, on peut citer les travaux de Synapse, pour le français notamment, dont le système repose sur les bases de connaissances et processus afférents qu'ils ont développés (Laurent *et al.*, 2005) et d'IBM avec Watson (Ferrucci *et al.*, 2010) sur l'anglais, qui met en œuvre de multiples stratégies et ressources. L'accent est mis sur l'amélioration de la résolution de sous-tâches, et notamment l'implication textuelle qui nous intéresse dans cet article. Les problèmes difficiles sont dûs à la variabilité linguistique et l'inférence, alors que le problème de l'ambiguïté ne se pose pas de manière cruciale : en effet les contextes formés par la question et le passage semblent suffisants pour résoudre les ambiguïtés, et cette question n'est pas abordée dans les travaux. Généralement, la résolution de l'appariement ne repose pas sur une mise en correspondance des constituants avec une ressource unique, et l'exploitation de diverses ressources permet d'alimenter des modèles d'apprentissage afin qu'ils prennent globalement la bonne décision. Dans ce cadre, les travaux mettent l'accent sur la définition de modèles permettant une intégration de connaissances diverses et une meilleure caractérisation des phénomènes à traiter.

3.4.2. *Techniques pour l'implication textuelle*

La majeure partie des systèmes de QR repose aujourd'hui sur des modèles discriminants, une classe de modèles d'apprentissage automatique qui modélise la dépendance à une entrée connue d'une sortie inconnue. Les algorithmes les plus utilisés sont sans doute la régression logistique (ou linéaire), les machines à vecteurs de support (SVM) et les réseaux de neurones. La régression peut être considérée comme l'algorithme d'apprentissage le plus intuitif et simple à implémenter, mais elle reste cependant très utile pour démontrer les bénéfices d'un choix de caractéristiques particulier (Yih *et al.*, 2013). Des travaux sont même capables de prédiction structurée en implémentant une régression logistique apprenant à la fois des alignements latents et la décision au problème de classification. Les SVM peuvent être utilisés exactement de la même façon, mais leur réel intérêt se trouve dans la possibilité de concevoir d'autres fonctions noyaux, adaptées au problème et aux données : on ne décide plus de ce qu'est *une* donnée, mais de ce que *deux* éléments ont en rapport. Les réseaux de neurones ont fait résurgence récemment, derrière l'appellation de *deep learning*, et permettent d'élaborer une architecture non linéaire complexe pour apprendre des représentations vectorielles adaptées à la tâche. Leur efficacité a été largement démontrée dans d'autres domaines, comme la description d'image (Socher *et al.*, 2014), mais reste encore à prouver sur des tâches de sémantique textuelle complexes.

3.4.3. *Connaissances modélisées : syntaxe* vs *sémantique*

Les modèles exploitant les représentations syntaxiques obtiennent de meilleurs résultats que les modèles « sac de mots », qui constituent la baseline. Cependant, ils

sont appris pour des questions dont la réponse est énoncée et justifiée en une phrase. De plus, il n'est pas clair de savoir quelles variations syntaxiques sont résolues. On peut supposer que ces modèles permettent d'apparier des syntagmes comportant des variations locales, qui sont les plus courantes, et des variations résultant d'erreurs d'analyse. Apprendre l'ensemble des variations nécessiterait sans doute de plus grands corpus. Les travaux de Yih *et al.* (2013) [13] ont tout particulièrement étudié comment réduire la distance sémantique de deux énoncés en intégrant différents types de relations sémantiques calculées à partir de multiples ressources, existantes ou apprises sur corpus. Ils montrent qu'un modèle de classification simple n'intégrant aucune information syntaxique permet d'obtenir les meilleurs résultats en sélection de phrases par rapport à un modèle intégrant un apprentissage de la structure latente d'alignement. La question reste donc ouverte quant au rôle joué par la syntaxe pour la sélection de passages réponses, en présence de connaissances sémantiques diverses.

3.4.4. *Types d'inférences évaluées*

Jusqu'à récemment, les phénomènes d'inférences évalués relevaient surtout de la similarité et du voisinage lexical, qui peuvent nécessiter des techniques et ressources certes avancées, mais rarement leur utilisation structurée, voire un raisonnement en plusieurs étapes. Des contributions récentes tentent de définir des inférences élémentaires nécessitant chacune un traitement particulier, comme un ensemble de tâches jouets (Weston *et al.*, 2015), ou les défis de Winograd (Levesque *et al.*, 2012). Par ailleurs, Gleize et Grau (2014) ont montré, par une annotation de données s'appuyant sur une hiérarchie de classes d'inférences, que les tâches évaluant la compréhension automatique de textes nécessitent de résoudre des phénomènes d'inférence plus complexes. L'étude approfondie de l'ensemble des phénomènes à prendre en compte passe par la construction de corpus représentatifs de ceux-ci. Afin de comparer les apports des différents modèles, qui sont généralement complexes et mettent en œuvre des ressources diverses, il faudrait que l'on puisse les évaluer plus finement que sur le résultat final de la tâche, ce que ne permettent pas les jeux de données actuels.

Après avoir exposé les systèmes recherchant les réponses dans les textes, nous présentons la recherche de réponses dans des bases de connaissances.

4. Recherche de réponses dans une base de connaissances

Lorsque la source d'information est composée d'une ou plusieurs bases de connaissances, la recherche d'une réponse diffère puisque l'information est déjà structurée, et donc la source d'information n'a pas besoin d'être analysée. Cependant, cette structuration existante contraint la représentation de la question qui doit être obtenue, puisqu'elle doit correspondre à celle de la base de connaissances. La recherche d'une réponse est alors ramenée à la construction d'une représentation de la question cohé-

13. `http://research.microsoft.com/pubs/192357/QA-SentSel-Updated.pdf` pour une version mise à jour.

rente avec la base de connaissances considérée, qui pourra être traduite en une requête en langage formel.

La construction de cette représentation nécessite d'identifier les ressources (entités et relations) de la base de connaissances dont il est question dans la question, et de produire une représentation de la question permettant d'obtenir la réponse. Le problème du choix du *modèle* pour représenter la question peut être approchée avec différents types de représentations intermédiaires. L'apprentissage de ces représentations requérant des données annotées manuellement, des travaux récents tentent de réduire la supervision en utilisant des représentations sémantiques latentes (Berant *et al.*, 2013). L'identification des *données* c'est-à-dire des ressources de la base de connaissances présentes dans la question nécessite de faire le lien entre le texte en langage naturel de la question et les triplets de la base de connaissances.

Afin de trouver la réponse parmi les ressources de la base de connaissances, plusieurs types de méthodes ont été proposés, qui se distinguent les uns des autres par plusieurs caractéristiques : certaines méthodes sont dirigées par l'analyse de la question, d'autres par la structure de la base de connaissances ; la représentation de la question peut être explicite, ou implicite ; les ambiguïtés peuvent être gérées à différentes étapes du processus ; les méthodes sont plus ou moins dépendantes de la base de connaissances utilisée ; les méthodes nécessitent plus ou moins de supervision.

L'identification des mentions des ressources étant des étapes communes aux différentes approches, nous les présentons d'abord, puis nous aborderons les représentations des questions, et enfin les processus d'interrogation des bases de connaissances.

4.1. *Difficultés de la mise en relation de la question avec la base de connaissances*

Afin d'identifier les entités et les relations, différents types de difficultés peuvent se poser, du fait du passage du langage naturel de la question à une représentation formelle. Dans les bases de connaissances, les entités et les relations possèdent en effet des *labels*, c'est-à-dire des expressions en langage naturel associées à la ressource, mais qui comprennent peu de variations. Ainsi, l'entité correspondant à la FIFA dans Wikidata a peu de labels, ce qui peut compliquer la reconnaissance d'une mention de l'entité, par exemple sous la forme « Fédération internationale de football association ». Plus généralement, il n'y a pas de correspondance unique entre une expression de la langue et une relation ou entité d'une base de connaissances du fait de l'ambiguïté du langage naturel. Une expression peut correspondre à plusieurs entités ou relations : ainsi, dans les deux questions « Où se situe le Louvre ? » et « Où se situe la glotte ? », l'expression « Où se situe » se traduira dans le premier cas par la relation *city* et dans le second par une relation d'inclusion. La formulation des relations peut en outre être implicite : ainsi, les questions « Citez un film de Spielberg. » et « Citez un livre de J. K. Rowling. » se traduiront par des relations *movie* et *author of*. Inversement, une ressource pourra se traduire par plusieurs formulations différentes : la relation *spouse* peut être utilisée pour répondre aux questions « Qui est la femme

de... ? », « Qui est l'époux de... ? », « À qui est marié... ? », « Qui ... a-t-il épousé ? » etc.

Les bases de connaissances n'étant pas nécessairement complètes ni cohérentes, d'autres problèmes peuvent se poser ; notamment une même relation sémantique peut avoir plusieurs instances différentes dans la base. Ainsi, les enfants peuvent être indiqués dans DBPedia par la relation *descendance* ou *enfants*. En outre, la granularité peut varier selon le type d'information considéré : certaines relations correspondent à une information très précise, d'autres sont beaucoup plus larges.

4.2. *Identification des entités*

L'une des entités de la question correspond à la réponse ; sa mention est généralement déterminée par une méthode basique, ou reprenant les classifications du cadre textuel. Le type de la réponse est également déterminé à cette étape.

L'identification des autres entités consiste à distinguer les termes de la question qui correspondent à des entités de la base (*annotation*), et à les associer aux entités correspondantes (*désambiguïsation*). Cette identification peut être effectuée avec un simple appariement de chaînes de caractères (Berant *et al.*, 2013), ou en calculant une similarité textuelle : Unger *et al.* (2012) utilisent les synonymes de WordNet pour obtenir des variantes des labels, puis appliquent des mesures de similarité textuelle afin d'annoter les entités. Bast et Haussmann (2015) prennent comme source de variation les données de CrossWiki, qui fournissent des variantes des entités de Wikipédia. L'annotation des entités peut également être réalisée par un outil spécifique, généralement fondé sur un apprentissage (supervisé ou non) : Zou *et al.* (2014) utilisent l'outil développé par Zhang *et al.* (2010) ; Hakimov *et al.* (2013) celui développé par Hakimov *et al.* (2012) ; Beaumont *et al.* (2015) et Park *et al.* (2015) l'outil Spotlight, créé pour DBPedia. Shekarpour *et al.* (2013) ont développé un outil de désambiguïsation fondé sur des modèles de Markov cachés. Enfin, certains systèmes apprennent à identifier les entités et relations avec le même processus, et sont donc présentés ci-dessous.

4.3. *Identification des relations*

L'identification des relations est au cœur du processus de réponse aux questions puisque les relations constituent l'unité de connaissance dans les bases. Cette identification nécessite de gérer les nombreuses variations possibles entre le label de la relation et ses mentions dans le texte, ainsi que de repérer les arguments de la relation.

Un premier type d'approche consiste à s'appuyer sur des lexiques pour reconnaître les mentions des relations, qui seront ensuite intégrées dans la représentation de la question. De nombreux travaux utilisent ainsi une base de paraphrases pour reconnaître la relation à partir de sa mention. L'une de ces bases est PATTY (Nakashole *et al.*, 2012), contenant des paraphrases apprises sur corpus, qui est utilisée notamment par Hakimov *et al.* (2013) ou Zou *et al.* (2014). Unger *et al.* (2012) comparent

les termes de la question à une autre base de patrons, extraits par l'outil BOA (Gerber et Ngomo, 2011). Berant *et al.* (2013) apprennent un lexique leur permettant d'aligner les termes de la question à des prédicats logiques. Fader *et al.* (2013) apprennent également un lexique permettant de relier des termes de la question à des relations de la base de connaissances ou à des patrons de requêtes (ainsi qu'à des entités).

Le second type d'approche repose sur le calcul d'une similarité sémantique entre des représentations continues de la question et de la base de connaissances. Yih *et al.* (2014) séparent ainsi les questions composées d'une seule relation en deux : la mention d'entité et le patron de relation, et calculent ensuite une similarité sémantique entre le patron de relation et les relations de la base de connaissances *via* un réseau de neurones convolutionnel. Bordes *et al.* (2014) utilisent un corpus de paraphrases de questions, extrait de WikiAnswers pour apprendre un score de similarité sémantique.

4.4. *Représentation de la question*

L'objectif étant de trouver la réponse dans une base de connaissances structurée, de nombreux systèmes visent à construire une représentation se rapprochant d'une requête par l'intermédiaire de représentations intermédiaires : graphes sémantiques (Yao et Van Durme, 2014 ; Zou *et al.*, 2014), patrons de requêtes (Unger *et al.*, 2012)... La construction de ces représentations se fonde généralement sur une analyse linguistique, au minimum syntaxique (Yao et Van Durme, 2014 ; Zou *et al.*, 2014), voire sémantique (Unger *et al.* (2012) utilisent par exemple des structures DRS). Yao et Van Durme (2014) partent ainsi du graphe de dépendances de la question, annotent certains nœuds comme l'interrogatif ou les entités nommées, et suppriment certaines dépendances en fonction de leur type pour obtenir un graphe de la question.

4.5. *Interrogation de la base de connaissances*

Lorsque la question n'est pas limitée à une seule relation, il convient ensuite de construire une représentation complète de la question à partir des entités et relations identifiées, et des éventuels opérateurs nécessaires à la requête. Cette construction de la requête peut être effectuée à partir de patrons prédéterminés. Unger *et al.* (2012) partent par exemple de la question, et commencent par appliquer des règles pour générer des patrons SPARQL génériques, qui seront ensuite instanciés en rapprochant les expressions de la question des ressources de la base de connaissances. Leur système obtient une f-mesure de 0,62 sur le sous-ensemble de questions de QALD utilisant les bases de connaissances intégrées dans le système. Les patrons étant prédéterminés, l'un des inconvénients de cette méthode est son manque de flexibilité.

La requête peut également être déduite de la représentation intermédiaire, souvent sous forme de graphe, approche choisie notamment par Zou *et al.* (2014), Beaumont *et al.* (2015) ou Shekarpour *et al.* (2013). Zou *et al.* (2014) adoptent par exemple une approche en deux étapes : l'analyse de la question, qui conduit à la construction

de graphes sémantiques de la question (les ambiguïtés sont conservées) et l'évaluation des requêtes, où des sous-graphes de la base de connaissances sont extraits en fonction de leur similarité avec le graphe de la question. Leur système obtient une f-mesure de 0,4 sur le jeu de questions de test de QALD3. Ce type d'approche ne prend pas en compte les questions nécessitant l'utilisation d'un opérateur comme « Who is the youngest player in the Premier League ? ». Yao et Van Durme (2014) ont une approche moins séquentielle, ne nécessitant pas de représentation intermédiaire complète de la question : plutôt que de générer une requête, ils extraient une vue de la base de connaissances constituée d'un sous-graphe autour des entités de la question, puis apprennent le processus d'extraction de la réponse dans ce sous-graphe en combinant des attributs de la question et du sous-graphe.

4.6. *Conclusion sur la recherche de réponses dans des bases de connaissances*

La recherche de réponses dans des bases de connaissances s'est principalement focalisée sur l'identification des relations dans les questions. Le problème majeur est la désambiguïsation des mentions de relation, qui peuvent correspondre à plusieurs relations différentes dans la base. Des lexiques appris et le typage des arguments de la relation peuvent être utilisés, ainsi qu'une mesure de similarité sémantique entre une représentation de la question et les triplets de la base. La construction d'une requête comprenant plus d'un triplet est généralement fondée sur l'analyse syntaxique de la question. Cependant cet aspect de la réponse aux questions est peu évalué, notamment du fait du manque de corpus suffisamment représentatifs : les corpus généralement utilisés dans les travaux cités sont principalement composés de questions dont la réponse peut être trouvée avec une seule relation. Les questions des évaluations QALD comprennent des questions dont les requêtes pour rechercher la réponse sont composées de plusieurs triplets ainsi que d'opérateurs, mais leur nombre est relativement limité.

5. Hybridation pour la recherche de réponses

Les textes et bases de connaissances ne contenant pas nécessairement les mêmes informations, plusieurs travaux ont cherché à exploiter ces deux types de ressources.

5.1. *Travaux existants*

Une première possibilité d'hybridation des deux types de systèmes consiste à rechercher des réponses dans des textes et dans des bases de connaissances en parallèle, puis à fusionner les réponses obtenues dans chacune des ressources (Hildebrandt *et al.*, 2004 ; Cucerzan et Agichtein, 2005). Clarke *et al.* (2002) commencent par rechercher une réponse dans des données structurées, puis, si aucune réponse n'y est trouvée, recherchent des informations dans des textes. La recherche de réponses se fait cependant dans une ressource d'un seul type.

Katz *et al.* (2005) décomposent les questions et ressources, structurées ou semi-structurées, afin d'obtenir une représentation sémantique commune, leur permettant de rechercher la réponse dans les deux types de ressources. Cependant, la robustesse de cette méthode n'est pas discutée, et le système n'est pas évalué.

Plus récemment, Yahya *et al.* (2013) ont mis en œuvre des techniques d'extension et de relaxation de requêtes afin de rechercher des informations dans des descriptions textuelles associées aux triplets : la première technique identifie les termes de la question qui ne sont pas reliés à un triplet dans la requête générée, et les ajoute comme mots-clés ; la deuxième identifie les triplets ou ensembles de triplets de la requête qui n'ont pas de correspondance dans la base, et les termes correspondant aux triplets menant à une absence de résultat sont ajoutés comme mots-clés ; enfin, la troisième technique, utilisée en dernier recours, consiste à considérer une simple requête pour vérifier le type de réponse attendu trouvé par l'analyse de la question, et ajouter les autres termes comme mots-clés associés. Les réponses ainsi obtenues sont ensuite réordonnées en fonction des entités saillantes et des mots-clés pertinents. L'utilisation des descriptions textuelles permet de passer d'un MRR de 0,53 à 0,72 sur le jeu de questions factuelles de QALD2, et d'obtenir une f-mesure de 0,68 (contre 0,74 pour le meilleur système participant). Usbeck et Ngomo (2015) ont une approche assez voisine, fondée sur la génération de requêtes hybrides : si aucune ressource n'est associée à un nœud du graphe de la question, ce nœud est ajouté à la requête sous forme textuelle.

Park *et al.* (2015) au contraire, recherchent d'abord des informations correspondant à la décomposition d'une question dans des textes annotés en entités de la base de connaissances, et complètent la recherche textuelle par des requêtes SPARQL si la recherche textuelle est infructueuse. Ce système obtient une bonne précision mais traite peu de questions sur le jeu de données hybrides de QALD5.

Le système Watson qui a participé à Jeopardy ! (Chu-Carroll *et al.*, 2012) effectue également une recherche hybride, mais en sélectionnant les relations à rechercher dans les bases de connaissances plutôt que comme stratégie de secours. Il utilise des bases de connaissances, soit existantes comme DBPedia, soit propres au système : la ressource PRISMATIC a ainsi été construite à partir de relations syntaxiques ou sémantiques extraites de textes. Les relations contenues dans ces bases sont exploitées de deux façons : la recherche de réponses lorsque l'une des relations Freebase les plus fréquentes est détectée dans la question, et la vérification du type de la réponse.

5.2. *Hybridation de processus de recherche*

L'hybridation de la recherche de réponses vise à étendre la couverture d'un système, sans présupposition de la ressource qui contient la réponse, et en exploitant au mieux les deux types de ressources. La question se pose quant aux types de questions qui peuvent bénéficier de cette hybridation. Ce point n'est pas discuté dans les travaux existants, car la réponse est *a priori* recherchée dans une ressource, l'autre

venant comme support ou palliatif à la première. De fait, les deux types de ressources ne contiennent pas les mêmes types d'information et nous allons donc examiner différentes caractéristiques des questions sous l'angle d'une résolution hybride. Les questions sont des questions factuelles dans les deux types de ressources, mais toutes ne peuvent profiter des informations provenant d'une base de connaissances. Cependant, lorsque la question mentionne une entité, une hybridation sera possible. Nous avons donc annoté en entités nommées un corpus de questions sur des textes comportant 9 227 questions. De manière à évaluer le taux d'erreur, nous avons vérifié l'annotation sur un échantillon de 200 questions. 59 % des questions comportent au moins une entité et entrent donc de fait dans les cas d'hybridation. Celle-ci peut fournir des informations intermédiaires nécessaires à la résolution de la question ou fournir la réponse, et cela selon les types d'information cherchés :

– la réponse est une caractéristique de l'entité (relation directe) : « Qui est le mari de la fille de Bill Clinton ? » La réponse peut être cherchée dans l'une ou l'autre source ;

– la réponse porte sur un événement, soit un rôle ou le nom de l'événement : « Qui est l'assassin de Martin Luther King ? » La réponse proviendra de textes, et d'autant plus si l'événement fait intervenir des entités non connues ;

– une combinaison des deux, par exemple une relation directe d'une entité ayant un rôle dans un événement (composition de relations) « Où est né l'assassin de Martin Luther King ? » On pourra effectuer une recherche hybride ;

– la réponse est une instance de concept ou un concept : « Quel animal pond des œufs bleus ? » La réponse viendra du texte, avec par exemple « Les Collonca sont sans queue et pondent des œufs bleus. », et la vérification que les Collonca sont des animaux pourra être opérée sur l'une ou l'autre source ;

– la relation avec la réponse est contextuelle (opinion ou liée à un événement par exemple comme dans « Quel pays a acheté du pétrole à l'Irak durant l'embargo ? ») Cette dernière ne pourra être trouvée que dans le texte ;

– une définition : « Qu'est-ce qu'un atome ? » La réponse est dans le texte ;

– un résultat provenant d'un opérateur d'agrégation (comparaison, classement, comptage) : « Donnez les dix plus grandes entreprises françaises. » Les réponses peuvent provenir de textes, dès lors que l'information cherchée est explicite, mais elles seront plus aisément déduites d'une base de connaissances.

Un intérêt supplémentaire à hybrider vient du fait que la langue et le schéma de la base ne font pas état du même niveau de granularité. Un mot peut faire référence à un chemin dans une base de connaissances et rendre la recherche dans cette base très complexe, comme pour la question « Donnez le nom d'un cosmonaute. » où « cosmonaute » doit être apparié à « astronaute de nationalité russe ». Inversement, le texte peut ajouter des informations, *i.e.* des contraintes, permettant de choisir la bonne entité parmi les possibles, par exemple « Quelle chanson des Rolling Stones parle de la fin d'une histoire d'amour ? » (réponse : *Angie*).

Examinons maintenant à quelles conditions l'hybridation est possible. Ainsi que nous l'avons vu dans l'état de l'art, les difficultés pour relier une question à sa réponse sont liées i) à la variabilité de la langue pour exprimer une relation, que ce soit pour ses mentions ou la reconnaissance de ses arguments, ii) à la présence d'informations implicites et iii) à la construction du sens d'un énoncé. Nous proposons de décomposer le problème global de représentation du sens d'une question en sous-problèmes unitaires, *i.e.* l'instanciation de relations binaires, pour lesquels le choix de la ressource interrogée dépend de la capacité des processus à résoudre les problèmes d'appariement. On pourra étendre à la fois la couverture du système de QR et sa capacité à répondre. Une telle approche a été proposée sur le texte pour modéliser l'implication textuelle (Cabrio et Magnini, 2011). En QR sur du texte, la décomposition a été exploitée dans (Moriceau *et al.*, 2009 ; Moriceau et Tannier, 2010) pour rechercher des réponses en exploitant des passages ou des documents différents, dans (Kalyanpur *et al.*, 2011) pour valider ou augmenter le contexte de recherche de réponses, et dans (Hartrumpf *et al.*, 2009) pour un système exploitant une représentation sémantique profonde et de surface. Dans ce cadre, la formalisation de la recherche de réponses à des questions quelle que soit la source d'information interrogée, peut reposer sur un modèle unifié de représentation atomique de relation (cf. figure 4).

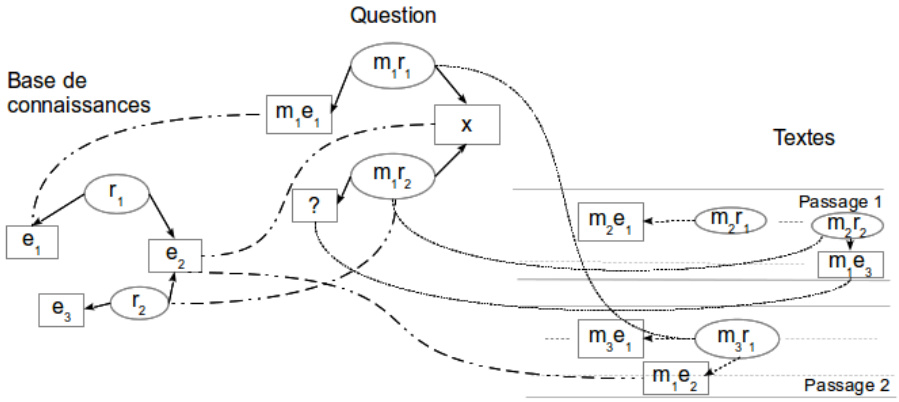

Figure 4. *Hybridation du processus de recherche*

Le sens global des questions et réponses est composé en suivant les différents liens d'appariement, vers l'une ou l'autre ressource. Afin de vérifier la cohérence de la représentation et savoir si les deux sources parlent des mêmes choses, il faut être en mesure de lier les informations trouvées dans le texte avec les informations issues de la base de connaissances. Cela nécessite un processus d'identification sémantique du texte vers la base de connaissances, et revient donc à normaliser relations et entités par celles qui sont présentes dans les bases interrogées, en conservant des informations textuelles supplémentaires. Ce processus est déjà présent lors de l'analyse de la question, et devra s'appliquer aussi aux passages de textes. Sur le plan des techniques de résolution, ce rapprochement des deux tâches peut aussi amener à poser les

problèmes de manière analogue, contrairement à ce qui est fait actuellement, et à les résoudre par les approches développées dans les deux domaines.

6. Bibliographie

Agirre E., Diab M., Cer D., Gonzalez-Agirre A., « Semeval-2012 task 6 : A pilot on semantic textual similarity », **SEM, International Workshop on Semantic Evaluation*, 2012.

Barr A., Feigenbaum E. A., « The Handbook of Artificial Intelligence. William Kaufmann », *Inc., Los Altos, CA*, 1981.

Bast H., Haussmann E., « More Accurate Question Answering on Freebase », *CIKM*, 2015.

Beaumont R., Grau B., Ligozat A.-L., « SemGraphQA@QALD5 : LIMSI participation at QALD5@CLEF », *Working Notes of CLEF 2015*, 2015.

Berant J., Chou A., Frostig R., Liang P., « Semantic Parsing on Freebase from Question-Answer Pairs. », *EMNLP*, p. 1533-1544, 2013.

Bernard G., Rosset S., Galibert O., Adda G., Bilinski E., « The LIMSI Participation in the QAst 2009 Track : Experimenting on Answer Scoring », *LNCS : Multilingual Information Access Evaluation I. Text Retrieval Experiments*, vol. 6241, p. 289-296, 2010.

Bollacker K., Evans C., Paritosh P., Sturge T., Taylor J., « Freebase : a collaboratively created graph database for structuring human knowledge », *SIGMOD*, 2008.

Bordes A., Weston J., Usunier N., « Open question answering with weakly supervised embedding models », *Machine Learning and Knowledge Discovery in Databases*, Springer, p. 165-180, 2014.

Bu F., Li H., Zhu X., « String re-writing kernel », *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, p. 449-458, 2012.

Cabrio E., Magnini B., « Towards component-based textual entailment », *Proceedings of the Ninth International Conference on Computational Semantics*, 2011.

Cai Q., Yates A., « Semantic parsing freebase : Towards open-domain semantic parsing », **SEM*, 2013.

Chen D., Manning C. D., « A fast and accurate dependency parser using neural networks », *EMNLP*, 2014.

Chu-Carroll J., Fan J., Boguraev B., Carmel D., Sheinwald D., Welty C., « Finding needles in the haystack : Search and candidate generation », *IBM Journal of Research and Development*, vol. 56, n° 3.4, p. 6-1, 2012.

Clarke C. L., Cormack G. V., Kemkes G., Laszlo M., Lynam T. R., Terra E. L., Tilker P. L., « Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002). », *TREC*, 2002.

Cucerzan S., Agichtein E., « Factoid Question Answering over Unstructured and Structured Web Content. », *TREC*, vol. 72, p. 90, 2005.

Cui H., Sun R., Li K., Kan M.-Y., Chua T.-S., « Question answering passage retrieval using dependency relations », *SIGIR*, 2005.

Dagan I., Glickman O., Magnini B., « The PASCAL recognising textual entailment challenge », *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, Springer, p. 177-190, 2006.

Dzikovska M. O., Nielsen R. D., Brew C., Leacock C., Giampiccolo D., Bentivogli L., Clark P., Dagan I., Dang H. T., SemEval-2013 task 7 : The joint student response analysis and 8th recognizing textual entailment challenge, Technical report, DTIC Document, 2013.

Fader A., Zettlemoyer L., Etzioni O., « Paraphrase-driven learning for open question answering », *ACL*, 2013.

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Jacquemin C., « Document selection refinement based on linguistic features for QALC, a question answering system », *RANLP*, 2001.

Ferrucci D., Brown E., Chu-Carroll J., Fan J., Gondek D., Kalyanpur A. A., Lally A., Murdock J. W., Nyberg E., Prager J. *et al.*, « Building Watson : An overview of the DeepQA project », *AI magazine*, vol. 31, n° 3, p. 59-79, 2010.

Ganitkevitch J., Van Durme B., Callison-Burch C., « PPDB : The Paraphrase Database », *Proceedings of NAACL-HLT*, 2013.

Gerber D., Ngomo A.-C. N., « Bootstrapping the linked data web », *1st Workshop on Web Scale Knowledge Extraction@ ISWC*, vol. 2011, 2011.

Gleize M., Grau B., « A hierarchical taxonomy for classifying hardness of inference tasks », *LREC*, 2014.

Gleize M., Grau B., « A Unified Kernel Approach for Learning Typed Sentence Rewritings », *ACL-IJCNLP*, 2015.

Glickman O., Applied textual entailment, PhD thesis, Bar Ilan University, 2006.

Glöckner I., Hartrumpf S., Leveling J., « Logical validation, answer merging and witness selection a study in multi-stream question answering », *RIAO*, 2007.

Grappy A., Grau B., Falco M.-H., Ligozat A.-L., Robba I., Vilnat A., « Selecting answers to questions from Web documents by a robust validation process », *WI*, 2011.

Hakimov S., Oto S. A., Dogdu E., « Named entity recognition and disambiguation using linked data and graph-based centrality scoring », *Proceedings of the 4th international workshop on semantic web information management*, 2012.

Hakimov S., Tunc H., Akimaliev M., Dogdu E., « Semantic question answering system over linked data using relational patterns », *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, 2013.

Hartrumpf S., « Extending knowledge and deepening linguistic processing for the question answering system InSicht », *Accessing Multilingual Information Repositories*, Springer, 2006.

Hartrumpf S., Glöckner I., Leveling J., « Efficient question answering with question decomposition and multiple answer streams », *Evaluating Systems for Multilingual and Multimodal Information Access*, Springer, p. 421-428, 2009.

Heilman M., Smith N. A., « Tree edit models for recognizing textual entailments, paraphrases, and answers to questions », *NAACL*, 2010.

Hildebrandt W., Katz B., Lin J. J., « Answering Definition Questions Using Multiple Knowledge Sources. », *HLT-NAACL*, p. 49-56, 2004.

Ittycheriah A., Franz M., Roukos S., « IBM's Statistical Question Answering System-TREC-10. », *TREC*, 2001.

Kalyanpur A., Patwardhan S., Boguraev B., Lally A., Chu-Carroll J., « Fact-based question decomposition for candidate answer re-ranking », *CIKM*, 2011.

Katz B., Borchardt G., Felshin S., « Syntactic and semantic decomposition strategies for question answering from multiple resources », *Proceedings of the AAAI 2005 workshop on inference for textual question answering*, p. 35-41, 2005.

Klein D., Manning C. D., « Accurate unlexicalized parsing », *ACL*, 2003.

Kouylekov M., Negri M., Magnini B., Coppola B., « Towards entailment-based question answering : ITC-irst at CLEF 2006 », *Evaluation of Multilingual and Multi-modal Information Retrieval*, Springer, p. 526-536, 2007.

Kozareva Z., Vasquez S., Montoyo A., « Adaptation of a Multi-learning Textual Entailment System to a Multilingual Validation Exercise », *Working Notes for the CLEF 2006 Workshop (AVE)*, 2006.

Laurent D., Séguéla P., Nègre S., « QRISTAL, système de Questions-Réponses », *TALN*, 2005.

Lehmann J., Isele R., Jakob M., Jentzsch A., Kontokostas D., Mendes P. N., Hellmann S., Morsey M., van Kleef P., Auer S. *et al.*, « DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia », *Semantic Web*, 2014.

Lehnert W., « Human and Computational Question Answering », *Cognitive Science*, vol. 1, n° 1, p. 47-73, 1977.

Levesque H. J., Davis E., Morgenstern L., « The Winograd schema challenge », *KR*, 2012.

Magnini B., Negri M., Prevete R., Tanev H., « Mining Knowledge from Repeated Co-Occurrences : DIOGENE at TREC 2002. », *TREC*, 2002.

Mihalcea R., Moldovan D. I., « Extended wordnet : Progress report », *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, 2001.

Miller G. A., « WordNet : a lexical database for English », *Communications of the ACM*, vol. 38, n° 11, p. 39-41, 1995.

Moldovan D., Clark C., Harabagiu S., Maiorano S., « Cogex : A logic prover for question answering », *NAACL*, 2003.

Moriceau V., Tannier X., « FIDJI : using syntax for validating answers in multiple documents », *Information retrieval*, vol. 13, n° 5, p. 507-533, 2010.

Moriceau V., Tannier X., Grau B., « Utilisation de la syntaxe pour valider les réponses à des questions par plusieurs documents », *CORIA*, 2009.

Moschitti A., Quarteroni S., Basili R., Manandhar S., « Exploiting syntactic and shallow semantic kernels for question answer classification », *ACL*, 2007.

Nakashole N., Weikum G., Suchanek F., « PATTY : a taxonomy of relational patterns with semantic types », *EMNLP-CoNLL*, 2012.

Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., Marsi E., « MaltParser : A language-independent system for data-driven dependency parsing », *Natural Language Engineering*, vol. 13, n° 02, p. 95-135, 2007.

Park S., Kwon S., Kim B., Lee G. G., « ISOFT at QALD-5 : Hybrid question answering system over linked data and text data », *Working Notes of CLEF 2015*, 2015.

Punyakanok V., Roth D., Yih W.-t., « Mapping dependencies trees : An application to question answering », *Proceedings of AI&Math 2004*, p. 1-10, 2004.

Punyakanok V., Roth D., Yih W.-t., « The importance of syntactic parsing and inference in semantic role labeling », *Computational Linguistics*, vol. 34, n° 2, p. 257-287, 2008.

Sachan M., Dubey A., Xing E. P., Richardson M., « Learning Answer-Entailing Structures for Machine Comprehension », *ACL*, 2015.

Sammons M., Vydiswaran V. V., Vieira T., Johri N., Chang M.-W., Goldwasser D., Srikumar V., Kundu G., Tu Y., Small K. *et al.*, « Relation alignment for textual entailment recognition », *Text Analysis Conference (TAC)*, 2009.

Shekarpour S., Ngonga Ngomo A.-C., Auer S., « Question answering on interlinked data », *Proceedings of the 22nd international conference on World Wide Web*, 2013.

Shen D., Lapata M., « Using Semantic Roles to Improve Question Answering. », *EMNLP-CoNLL*, p. 12-21, 2007.

Socher R., Karpathy A., Le Q. V., Manning C. D., Ng A. Y., « Grounded compositional semantics for finding and describing images with sentences », *Transactions of the Association for Computational Linguistics*, vol. 2, p. 207-218, 2014.

Suzuki J., Sasaki Y., Maeda E., « SVM answer selection for open-domain question answering », *COLING*, 2002.

Tatu M., Iles B., Slavick J., Novischi A., Moldovan D., « Cogex at the second recognizing textual entailment challenge », *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, p. 104-109, 2006.

Unger C., Bühmann L., Lehmann J., Ngonga Ngomo A.-C., Gerber D., Cimiano P., « Template-based question answering over RDF data », *WWW*, 2012.

Usbeck R., Ngomo A.-C. N., « HAWK@ QALD5–Trying to answer hybrid questions with various simple ranking techniques », *Working Notes of CLEF 2015*, 2015.

Volokh A., Neumann G., « 372 : Comparing the benefit of different dependency parsers for textual entailment using syntactic constraints only », *SemEval*, 2010.

Wang M., Manning C. D., « Probabilistic tree-edit models with structured latent variables for textual entailment and question answering », *COLING*, 2010.

Wang M., Smith N. A., Mitamura T., « What is the Jeopardy Model ? A Quasi-Synchronous Grammar for QA. », *EMNLP-CoNLL*, vol. 7, p. 22-32, 2007.

Weston J., Bordes A., Chopra S., Mikolov T., « Towards AI-complete question answering : a set of prerequisite toy tasks », *arXiv preprint arXiv :1502.05698*, 2015.

Yahya M., Berberich K., Elbassuoni S., Weikum G., « Robust question answering over the web of linked data », *CIKM*, 2013.

Yao X., Van Durme B., « Information extraction over structured data : Question answering with freebase », *ACL*, 2014.

Yao X., Van Durme B., Callison-Burch C., Clark P., « Semi-Markov Phrase-Based Monolingual Alignment. », *EMNLP*, p. 590-600, 2013.

Yih W.-T., Chang M.-W., Meek C., Pastusiak A., « Question Answering Using Enhanced Lexical Semantic Models », *ACL*, 2013.

Yih W.-t., He X., Meek C., « Semantic parsing for single-relation question answering », *ACL (Short Papers)*, 2014.

Zhang W., Su J., Tan C. L., Wang W. T., « Entity linking leveraging : automatically generated annotation », *COLING*, 2010.

Zou L., Huang R., Wang H., Yu J. X., He W., Zhao D., « Natural language question answering over RDF : a graph data driven approach », *SIGMOD*, 2014.