Typologie des questions à réponses multiples pour un système de question-réponse*

Mathieu-Henri Falco
LIMSI-CNRS, Université Paris-Sud, 91403 Orsay, France
falco@limsi.fr

RÉSUMÉ

L'évaluation des systèmes de question-réponse lors des campagnes repose généralement sur la validité d'une réponse individuelle supportée par un passage (question factuelle) ou d'un groupe de réponses toutes contenues dans un même passage (questions listes). Ce cadre évaluatif empêche donc de fournir un ensemble de plusieurs réponses individuelles et ne permet également pas de fournir des réponses provenant de documents différents. Ce recoupement inter-documents peut être necessaire pour construire une réponse composée de plusieurs éléments afin d'être le plus complet possible. De plus une grande majorité de questions formulées au singulier et semblant n'attendre qu'une seule réponse se trouve être des questions possédant plusieurs réponses correctes. Nous présentons ici une typologie des questions à réponses multiples ainsi qu'un aperçu sur les problèmes posés à un système de question-réponse par ce type de question.

ABSTRACT

Typology of Multiple Answer Questions for a Question-answering System

The evaluation campaigns of question-answering systems are generally based on the validity of an individual answer supported by a passage (for a factual question) or a group of answers coming all from a same supporting passage (for a list question). This framework does not allow the possibily to answer with a set of answers, nor with answers gathered from several documents. This cross-checking can be needed for building an answer composed of several elements in order to be as accurate as possible. Besides a large majority of questions with a singular form seems to be answered with a single answer whereas they can be satisfied with many. We present here a typology of questions with multiple answers and an overview of problems encountered by a question-answering system with this kind of questions.

MOTS-CLÉS : question-réponse, questions à réponses multiples, question liste.

KEYWORDS: question-answering, multiple answer questions, list question.

*Ces travaux ont été partiellement financés par OSEO dans le cadre du programme QUAERO.

1 Introduction

Les systèmes de question-réponse (*SQR*) ont pour but de fournir une réponse précise à une question formulée en langue naturelle par un utilisateur : ils peuvent travailler à partir de bases de données et/ou de collections de documents ; nous nous intéressons ici uniquement à ceux interrogeant un corpus de documents. Ces SQR combinent plusieurs domaines dont notamment la recherche d'information et le TAL à travers l'extraction d'informations. En effet, là où des moteurs de recherche renvoient des références de documents (avec éventuellement un extrait) suite à une requête sous forme de mots-clés, les SQR travaillent à partir d'une question en langue dont tous les mots ne sont pas forcément pertinents pour la recherche d'information, sélectionnent un ensemble de documents de la collection puis extraient la réponse précise de ces documents afin de la présenter à l'utilisateur (éventuellement accompagnée de l'extrait contenant cette réponse) .

Les SQR existants utilisent des approches variées, pouvant s'appliquer sur la totalité du système ou seulement certaines parties. Par exemple, des systèmes utilisent une représentation de la question et des documents logique (Moldovan *et al.*, 2007) ou discursive (Bos *et al.*, 2007a). La syntaxe peut être utilisée au niveau de l'extraction de la réponse : par exemple pour la fusion d'information multidocuments à l'aide de dépendances syntaxiques (Moriceau et Tannier, 2010), (Katz et Lin, 2003). Des heuristiques de distance (Fangtao *et al.*, 2008) ou un apprentissage automatique (Grappy, 2011) peuvent être utilisés pour la validation d'un candidat-réponse. Au final, les SQR se trouvent souvent bridés par le processus d'évaluation actuelle des campagnes à savoir fournir, pour chaque question, plusieurs réponses (généralement de trois à cinq) sous la forme d'une paire référence du document/réponse précise éventuellement accompagnée d 'un extrait du document d'où la réponse a été extraite (passage justificatif) comme pour Quaero 2010 (Quintard *et al.*, 2010) ou TREC-2007 (Dang *et al.*, 2007) (toutes les éditions et guidelines des campagnes CLEF[1] et TREC[2] sont en ligne).

Nous avons choisi de nous intéresser plus particulièrement aux questions que nous appellerons "questions à réponses multiples" comme par exemple les questions *Quels sont les sept astres du système solaire visibles à l'oeil nu ?* (le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter, Saturne) ou *Quand le Paris Saint-Germain a-t-il été sacré champion de France de football ?* (1986 et 1994 pour l'équipe professionnelle homme). Ces questions ne sont que peu ou pas évaluées lors des campagnes d'évaluation des SQR. Pourtant, elles présentent un intérêt tant au niveau de l'analyse de la question que de l'extraction des réponses et de leur présentation à l'utilisateur. En effet, un système doit être capable dans le meilleur des cas d'extraire une liste de réponses bien formée d'un document mais, le plus souvent, le système doit reconstituer une liste à partir d'éléments provenant de documents différents. Nous avons choisi de nous intéresser aux SQR qui interrogent le Web car cela nous permet de travailler en domaine ouvert et, étant donné le nombre important de documents, le travail de recoupement des réponses multi-documents s'avère indispensable.

Dans cet article, nous commençons donc par définir le terme **question à réponses multiples** (*question-ARM*) et présentons un état de l'art concernant le traitement de ce type de questions par les SQR ainsi que les élèments structuraux sources de réponses de type liste. Dans la section 3, nous présentons les observations constatées sur les données de deux campagnes d'évaluation proposant des questions-ARM. Les sections 4 et 5 présenteront respectivement notre corpus

[1] `http://nlp.uned.es/clef-qa/`
[2] `http://trec.nist.gov/data/qamain.html`

d'étude et les typologies que nous avons définies. Enfin, la section 6 présentera les perspectives envisagées.

2 Contexte et état de l'art

Une question à réponses multiples est une question pour laquelle plusieurs réponses peuvent être correctes. La forme la plus évidente de réponse à ce type de question est bien sûr la liste, par exemple :

question : *Quels sont les sept astres du système solaire visibles à l'oeil nu ?*
réponse : le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter, Saturne
passage : Les astres visibles à l'oeil nu, **le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter et Saturne** tiennent leur nom du monde romain.

Les éléments composant la liste de réponses peuvent bien sûr être déjà sous la forme d'une liste dans un document mais ils peuvent aussi être répartis dans un document ou même dans plusieurs documents, par exemple :

question : *Quand le Paris Saint-Germain a-t-il été sacré champion de France de football ?*
réponse : 1986 et 1994
document 1 : Le PSG champion de France **1986**, vu par son entraîneur, Gérard Houiller.
document 2 : Les positions resteront les mêmes durant tout le reste de la saison : le PSG sera sacré champion de France **1994**.

Nous présentons ici comment ces questions sont abordées par les SQR ainsi que les éléments structuraux qui permettent d'identifier les réponses dans les textes.

2.1 L'évaluation des SQR

L'évaluation des SQR peut se faire au niveau de la satisfaction utilisateur (point de vue applicatif et qualitatif) ou par l'intermédiaire d'une métrique (point de vue comparatif car quantitatif). Les campagnes d'évaluations de SQR ont pour but de jauger les performances des différentes approches et proposent pour cela un nombre de questions significatif pour les catégories les plus fréquentes : *factuelles*, *booléennes*, *définition*, *complexes*, *liste* et *nil* (questions n'ayant pas de réponses dans la collection de documents). Les systèmes doivent fournir plusieurs réponses pour chaque question (de trois à cinq généralement) et sont le plus souvent évalués grâce à la métrique du *MRR* (Mean Reciprocal Rank) qui favorise ainsi les SQR fournissant une réponse correcte dans les premiers rangs.

Il est très difficile de garantir qu'une unique réponse correcte puisse être obtenue à partir de la collection de documents disponible pour l'évaluation, ce qui serait peu intéressant d'ailleurs, et une évaluation humaine des réponses doit parfois avoir lieu pour juger la réponse ainsi que le passage justificatif.

Dans les campagnes traitant des *question-listes* (questions de type liste), pour indiquer qu'on

n'attend pas une réponse unique, une marque de pluriel est toujours présente mais le nombre de réponses attendues n'est pas toujours mentionné dans la question comme dans les exemples suivants : *Quelles sont les 4 localisations possibles des neuroblastomes ?* (EQueR, Quaero 2008, 2009, TREC 2001, 2002) ou *Quels sont les secteurs qui recrutent ?* (Quaero 2010, TREC 2003 à 2007). La métrique utilisée pour évaluer les réponses à ce type de questions est alors dans le premier cas la précision moyenne (nombre de réponses correctes/nombre de réponses attendues) et dans le second la F-mesure (en considérant l'ensemble des réponses jugées correctes par les évaluateurs.

Par l'utilisation du MRR, les campagnes analysant un triplet question/réponse/passage obligent donc les SQR à faire un choix d'au plus N réponses par question. Une réponse issue d'un recoupement d'informations entre plusieurs documents est donc difficile à justifier dans le cadre d'une campagne d'évaluation. De plus, la réponse et le passage doivent obligatoirement être du texte issu d'un document de la collection alors qu'il peut être parfois plus pertinent de renvoyer un élément structural (un tableau par exemple). Ces éléments structuraux sont très présents dans les documents Web mais, de toutes les campagnes évoquées jusqu'à présent, seule Quaero utilise une collection de documents Web et impose un format de réponse assez identique à celui des autres campagnes (Quintard *et al.*, 2010).

2.2 Le traitement des questions à réponses multiples par les SQR

Les questions-listes sont un cas particulier des questions-ARM : elles attendent comme réponse une liste d'items provenant d'une même entité (phrase ou document). Parmi les SQR ayant participé aux campagnes proposant des questions dont la réponse est de type liste, plusieurs ont adapté leur traitement de questions factuelles aux listes. Cette adaptation consiste à utiliser la liste ordonnée de leurs réponses trouvées dans la collection en renvoyant directement un top-N de leurs réponses, le nombre N pouvant être fixe (par exemple 5 pour (Chu-carroll *et al.*, 2004) et 20 pour (Wu *et al.*, 2003)) ou dépendant d'un nombre de réponses attendues présent dans la question (Bos *et al.*, 2007b) ou d'un seuil déterminé par le SQR selon son système d'ordonnancement (Kaisser et Becker, 2004) (Schlaefer *et al.*, 2007).

Les SQR ayant développé un traitement spécifique pour les listes ont notamment utilisé la détection de doublon pour éviter la redondance de candidat-réponse (Katz *et al.*, 2006) et certains utilisent en plus la réconciliation de référence à l'aide de ressources extérieures (Schlaefer *et al.*, 2007), (Dan I. Moldovan and et Bowden, 2007). À travers l'expansion de requête, la co-occurrence des candidats-réponses, au niveau de la phrase ou du document, sert notamment de critère de validation (Razmara et Kosseim, 2008) (Wang *et al.*, 2008) (Figueroa et Neumann, 2008).

La plupart de ces SQR utilisent donc des ressources extérieures comme des bases de données ou le web, or nous ne souhaitons pas en dépendre et seulement utiliser une collection de documents. De plus l'aspect multi-document n'est vu généralement qu'en phase de validation par la co-occurrence.

2.3 Les éléments structuraux

Nous nous sommes intéressés aux éléments structuraux que sont l'objet tableau et l'objet liste car nous nous attendons à ce qu'ils soient une source de réponses à des questions-ARM (nous ne nous intéressons qu'aux données textuelles et avons donc laissé de côté les images, figures, animations flash, etc.). Nous considérons ici le terme d'objet tableau comme une structure à au moins deux lignes et deux colonnes, et l'objet liste comme une constitution de plusieurs entités disposées horizontalement (énumération dans une phrase) ou verticalement (amorce se terminant par le symbole deux-points et un item par ligne par exemple).

Les listes ont été l'objet d'études approfondies du point de vue discursif afin de mieux cerner la structure d'un document (Ho-Dac, 2007). Les travaux de (Péry-Woodley, 2000), (Luc, 2001), (Bras *et al.*, 2008), (Laignelet, 2009), (Ho-Dac *et al.*, 2010) ont beaucoup traité de cette question et ont ainsi défini l'objet répondant au terme de "structure énumérative" comme étant composé d'une amorce (phrase introductrice), d'une énumération composée d'items (entité co-énumérée caractérisée par diverses marques typographiques, dispositionnelles, lexico-syntaxiques). Plusieurs travaux applicatifs se sont intéressés aux listes dans le cadre du peuplement d'ontologie (Laignelet *et al.*, 2011), les entités nommées (Jacquemin et Bush, 2000) et de l'analyse syntaxique. En effet, l'objet liste est par nature difficile à analyser syntaxiquement au sens où il peut utiliser la verticalité, une autre ponctuation (le point-virgule entre les items), une typographie assez libre (choix des puces) et créer des liens syntaxiques entre les items, l'amorce et la conclusion. Xerox a creusé dans cette direction à travers les travaux de (Aït-Mokhtar *et al.*, 2003) et (Gala, 2003).

Les tableaux ont été traités du point de vue HTML avec pour but de typer les cases, soit à des fins de visualisations ergonomiques, soit pour de l'extraction d'information. Deux types d'approches dominent : à bases de règles (Gatterbauer *et al.*, 2007), (Tajima et Ohnishi, 2008) et par apprentissage automatique sur un corpus annoté manuellement (Wang et Hu, 2002).

3 Premières observations sur des corpus de campagnes d'évaluation proposant des questions-ARM

Nous nous sommes intéressés dans un premier temps aux campagnes pour le français EQueR-Evalda 2004 (Ayache *et al.*, 2006) et QUAERO 2008 (Quintard *et al.*, 2010) qui comportaient des questions-listes et pour lesquelles nous avions accès aux collections (voir tableau 1). La campagne d'évaluation EQueR a proposé deux tâches : une générique sur une collection hétérogène de textes journalistiques (désignée ici par *Eq-Jour*) et une spécifique sur une collection de textes médicaux (désignée par *Eq-Méd*).

3.1 Collecte des données

Dans un premier temps, nous sommes partis des questions typées par les évaluateurs des campagnes et avons étudié les questions-listes. Ces questions portaient toutes une marque de pluriel sous forme de nombre de réponses attendu et se formulaient syntaxiquement sous quatre patrons (voir tableau 2).

	Eq-Méd	Eq-Jour	Quaero
Domaine	médical	presse, politique	ouvert
Format des documents	texte	texte	HTML
Nombre de documents	5 623	557 300	499 736
Taille de la collection	0,135 Go	1,5 Go	5 Go
Nombre de questions-listes	25	30	29

TAB. 1 – Caractéristiques des corpus EQueR et Quaero

	Eq-Méd	Eq-Jour	Quaero
Citez X	13	5	14
Quels sont les X ?	12	22	15
Qui sont les X ?	0	2	0
Comment se prénommaient les X ?	0	1	0
Nombre de questions-listes	25	30	29

TAB. 2 – Nombre de questions-listes par forme syntaxique (X est le nombre de réponses attendu).

Nous avons ensuite effectué une première étude des documents contenant des réponses correctes, documents fournis par les organisateurs des campagnes sous forme d'un fichier de référence. Nous considérons ici qu'une réponse est correcte si elle répond à la question (validation humaine), même s'il existe des réponses correctes plus pertinentes (au sens de plus récentes par exemple, ou bien satisfaisant plus l'utilisateur dans un cadre applicatif), et même si la question attendait plusieurs réponses de façon explicite (nombre déterminé dans la question). Nous utiliserons le terme *réponse-liste* pour désigner le groupe de réponses correctes à une question en attendant un nombre déterminé.

Nous avons utilisé le moteur de recherche Lucene[3] pour rechercher les documents contenant au moins une réponse aux questions-listes puisque les réponses de fichier de référence n'étaient pas forcément exhaustives. Les requêtes ont été formulées soit à partir des termes de la question jugés importants, soit à partir des réponses des fichiers références. Nous avons étudié manuellement jusqu'à 50 extraits de documents par question puis les documents entiers si les snippets étaient pertinents. Les requêtes ont ensuite été reformulées à l'aide de synonymes pour les termes de la question et des réponses nouvellement trouvées afin d'augmenter le nombre de passages-réponses. Nous avons arrêté la collection quand nous n'observions plus de nouvelle réponse ou de nouveaux phénomènes.

3.2 Étude des couples questions-ARM/réponses

L'étude des réponses aux questions-listes a révélé un nombre important de passages-réponses pour les trois collections (voir tableau 3). Les résultats montrent que la forme préférentielle d'une réponse à une question-liste dans ces campagnes est majoritairement la phrase comme par exemple :

[3] `http://lucene.apache.org/core/`

	Eq-Méd	Eq-Jour	Quaero
Nombre de questions-listes	25	30	29
Nombre de passages-réponses	56	112	122
Nombre moyen de passages-réponses par question	2,33	3,73	4,21
Passage sous forme de phrase	**30 (51,85 %)**	**73 (69,67 %)**	**85 (70,25 %)**
Passage sous forme de paragraphe	11 (20,37 %)	37 (33,04 %)	19 (15,57 %)
Passage sous forme de liste	15 (27,78 %)	2 (1,79 %)	17 (13,93 %)
Passage sous forme de tableau	0 (0 %)	0 (0 %)	1 (0,82 %)

TAB. 3 – Forme du passage-réponse.

	Eq-Méd	Eq-Jour	Quaero
Nombre de questions-listes	24	30	29
Nombre de questions avec plusieurs passages-réponses dans un même document	8 (33,33 %)	4 (13,33 %)	12 (41,38 %)
Nombre de questions où un document contient une réponse-liste	12 (50 %)	18 (60 %)	22 (75,86 %)
Nombre de questions où la réponse-liste est obligatoirement inter-document	**6 (25 %)**	**0 (0 %)**	**0 (0 %)**
Nombre de passages-réponses	56	112	122

TAB. 4 – Répartition des réponses dans les documents.

question : *Quels sont les 7 astres du système solaire visibles à l'oeil nu ?*
passage-réponse : *Les astres visibles à l'oeil nu, le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter et Saturne tiennent leur nom du monde romain.*

Cette répartition centrée sur un bloc de texte continu (phrase, paragraphe, liste) contenant toutes les réponses est due aux choix des organisateurs de la campagne. Les questions-listes générées pour Quaero l'étaient notamment sur la base d'un document devant contenir tous les éléments permettant d'y répondre.

Nous avons ensuite étudié la répartition des passages-réponses dans les documents (voir tableau 4). Il en a résulté une confirmation d'une redondance des réponses inter-documents et également intra-document. La redondance inter-documents était totale au sens où chaque document contenant une réponse correcte contenait aussi la réponse-liste : seul un quart des questions de Eq-Méd nécessitait au moins deux documents de Eq-Méd pour pouvoir composer l'ensemble des réponses attendues (il n'existait pas de document unique contenant le nombre de réponse attendu pour 25 % de ces questions de Eq-Méd).

Devant le peu de questions soulevant une nécessité de traitement inter-document dans ces collections, nous avons donc décidé de constituer un autre corpus d'étude pour les questions-ARM.

4 Corpus d'étude pour les questions-ARM

Afin d'étudier en détail la forme des réponses dans le but de mieux les extraire automatiquement, nous avons constitué un corpus d'étude pour les questions-ARM en générant d'abord des questions-ARM puis en récupérant des documents permettant d'y répondre.

4.1 Constitution et caractéristiques du corpus

Nous avons d'abord repris sept questions listes existantes dans EQueR et Quaero en supprimant le nombre de réponses attendu (*Qui sont les ~~huit~~ personnages de "Disney Princess" ?*) ou en changeant des termes (*Quels pays étaient candidats à l'organisation de la coupe du monde 2018 ?* au lieu de *2006*). Nous avons ensuite imaginé des questions propices aux réponses multiples : par exemple, *Qui a incarné Batman ?*, *Quand la France a-t-elle perdu son triple-A ?*. En utilisant trois moteurs de recherche sur Internet (Exalead[4], Bing[5] et Google[6]), nous avons collecté des documents contenant au moins une réponse correcte. Ainsi un document peut ne contenir qu'un seul pays candidat à l'organisation de la coupe du monde 2018 ou qu'un seul nom d'acteur ayant incarné Batman et non pas forcément la réponse-liste. Si ce document contenait une ou des réponses dans un tableau ou une liste, il était ajouté au corpus au même titre que les autres documents.

Le corpus d'étude se compose actuellement de cent questions ayant été générées manuellement sur des thématiques variées (sport, santé, politique, culture, économie, informatique) et sous plusieurs types. Les informations sont présentées de la façon suivante : type de la question (nombre de questions dans le corpus) (nombre de questions necessitant un traitement inter-document pour répondre pertinemment) : exemple.
– factuelle (61) (11) : *Quand la France a-t-elle perdue son triple-A ?* ;
– liste (17) (2) : *Quels pays étaient candidats à l'organisation de la coupe du monde 2018 ?* ;
– complexe (10) (3) : *Comment a évolué la croissance française en 2011 ?* ;
– booléenne (8) (3) : *Pluton est-elle une planète ?* ;
– définition (4) (0) : *Qu'est-ce que la croissance ?*.
Pour ces questions, nous avons récupéré 232 fichiers au format HTML, chacun d'entre eux contenant donc au moins une réponse correcte. Au total, seules 19 questions nécessitent obligatoirement un traitement inter-document pour composer la réponse. Cette basse proportion s'explique notamment par le fait que quelques pages très pertinentes (Wikipédia notamment) contenaient effectivement toutes les réponses. Nous avons décidé de les garder car les réponses étaient réparties sur l'ensemble du document (souvent de très grande taille). De plus leur identification nous servira de baseline pour mesurer les résultats sans traitement inter-document.

4.2 Observations

L'observation des passages-réponses (voir tableau 5) a d'abord montré des problèmes récurrents des SQR pour lesquels il existe déjà une base de travaux s'y intéressant, à savoir la résolution d'anaphore, la réconciliation de référence, le type métaphorique de la formulation de réponse

[4] http://www.exalead.fr/search/
[5] http://www.bing.com/
[6] http://www.google.fr

(*Le triple A, c'est une ligne Maginot.*), le besoin de contexte, les faux candidats (*en France nous avons le quintuple A (amicale des amateurs d'andouillettes authentique*)), pour ne citer qu'eux.

Le recensement précis a montré plusieurs phénomènes dont les plus émergents sont :
– les réponses se trouvant dans des tableaux de données ce qui confirme le besoin de savoir les analyser ;
– la présence d'informations incertaines (par exemple, des rumeurs ou avec l'usage du conditionnel) ;
– les réponses sont réparties dans des chronologies narratives (document retraçant chronologiquement un thème) ;
– le recoupement d'informations réparties dans plusieurs documents.

Nombre OCC	PHÉNOMÈNE
82	critère variant
72	formulation de la réponse
53	ancre référentielle à chercher
46	faux candidats
21	rumeur
20	chronologie narrative
18	tableau de données
17	terminologie dans la question
13	indice d'expansion de requête
12	beaucoup de candidats-réponses du type attendu dans un court passage textuel
12	besoin de contexte
11	type métaphorique de la réponse
10	terminologie dans la réponse

TAB. 5 – Occurrences des phénomènes (non mutuellement exclusifs) recensés les plus fréquents.

Nous présentons ici les 3 phénomènes auxquels nous avons choisi de nous intéresser par la suite car ce sont les plus fréquents dans notre corpus.

Le phénomène le plus fréquent est la variation des réponses selon certains critères : un critère de précision (que nous appellerons **critère variant**) d'un élément permet de créer plusieurs réponses correctes. Ici, la note souveraine de la France dépend de l'agence de notation :

question : *Quelle est la note de la France sur les marchés financiers ?*
passage-réponse 1 : *L'agence de notation américaine Egan-Jones a abaissé aujourd'hui la note attribuée à la dette de la France à "A", cinq crans en dessous du "triple A" des trois grandes du secteur, Standard and Poor's, Moody's et Fitch.(—). La France avait perdu son "AAA" chez cette agence en juillet.*
passage-réponse 2 : *"Moody's a maintenu le triple A de la France, la meilleure note possible", annonçait le matin une dépêche AFP, aussitôt reprise par une partie de la presse française.*
passage-réponse 3 : *Peu après 16 heures, ce vendredi, une source gouvernementale a indiqué que l'agence de notation financière Standard & Poor's avait bel et bien décidé de dégrader la France en lui retirant sa note d'excellence triple A.*

Le problème de **la formulation de la réponse** est un aussi problème habituel des SQR : la réponse, par la synonymie ou la paraphrase, peut prendre plusieurs formes :

question : *Qui a incarné Batman ?*
passage-réponse 1 : *Après avoir usé **Michael Keaton**, **Val Kilmer** et **George Clooney** dans le rôle de Batman, les spéculations sur le prochain vengeur masqué de Gotham City se poursuivent.*
passage-réponse 2 : *Le réalisateur chinois Zhang Yimou a choisi pour son prochain film l'acteur britannique **Christian Bale**, qui a incarné Batman, pour jouer le rôle d'un prêtre héroïque durant le sac de Nankin par les troupes japonaises en 1937.*

L'ancre référentielle est le phénomène nécessitant un besoin de rattachement à une date précise. En effet, le temps est un critère variant et les réponses correctes ne le sont parfois que par rapport à un moment temporel précis. Par exemple dans les trois passages-réponses suivants, les réponses nécessitent de trouver la date absolue à partir des indices temporels relatifs (en gras) pour pouvoir être validées :

question : *Quand la France a-t-elle perdu son triple-A ?*
passage-réponse 1 : *L'agence de notation américaine Egan-Jones a abaissé aujourd'hui la note attribuée à la dette de la France à "A", cinq crans en dessous du "triple A" des trois grandes du secteur, Standard and Poor's, Moody's et Fitch.(—).La France avait perdu son "AAA" chez cette agence **en juillet**.*
passage-réponse 2 : *"Moody's a maintenu le triple A de la France, la meilleure note possible", annonçait **le matin** une dépêche AFP, aussitôt reprise par une partie de la presse française.*
passage-réponse 3 : ***Peu après 16 heures, ce vendredi**, une source gouvernementale a indiqué que l'agence de notation financière Standard & Poor's avait bel et bien décidé de dégrader la France en lui retirant sa note d'excellence triple A.*

5 Expérimentation dans un cadre classique

Nous avons soumis au SQR FIDJI (Moriceau et Tannier, 2010) les cent questions de notre corpus afin d'analyser son comportement prévu pour une campagne d'évaluation classique. L'étude des résultats nous a permis dans un premier temps de mieux catégoriser les questions-ARM afin d'en dresser une typologie et dans un deuxième temps de mieux cibler les difficultés à résoudre pour pouvoir y répondre dans le futur.

5.1 Typologie des questions-ARM

L'étude avait révelé que 47 des 61 questions factuelles se révélaient être potentiellement des questions-ARM. Nous avons donc étudié les phénomènes composant ces questions-ARM en plus des questions-listes afin d'être en mesure de les typer lors de l'analyse des questions (figure 1). La marque du pluriel sur le focus de la question indique explicitement une question-ARM tandis que certains indices (notion temporelle, granularité du pronom *qui* et des adverbes interrogatif *où* et *quand*) peuvent potentiellement indiquer une question-ARM mais seules les réponses permettront au final de trancher. Le critère variant le plus fréquent (53,55 %) est le critère temporel mais il peut être plus général : les questions étant souvent courtes, le sens prototypique des concepts est fréquemment utilisé. Ainsi, parmi les exemples suivants de questions illustrant les phénomènes de

la typologie en figure 1, la question (6) pour un français passionné de football fait communément référence à la ligue des champions de football masculine et européenne alors qu'aucun de ces deux termes n'est présent :

– (identifiant en figure 1) Pourcentage sur les 100 questions du corpus : *Question* explication sur les réponses"
– **(1), 10 %** : *Quels ministères a occupé Alliot-Marie ?* "La Défense, l'Intérieur..." ;
– **(2) 7 %** : *Quels sont les pays de l'UE ?* "France, Finlande, Allemagne..." (27 pays en 2012) ;
– **(3) 1 %** : *Quelles sont les neuf planètes du système solaire ?* "Mercure, Vénus, Terre..." ;
– **(4) 33 %** : *Où/Quand/À qui Sarkozy a-t-il présenté ses voeux 2012 ?* "À Lille le 12 janvier aux fonctionnaire, À Lyon le 19 janvier au monde économique..." ;
– **(5) 11 %** : *Qui sont les Disney Princess ?* "Tiana a été ajoutée en 2009 à la collection, Raiponce en 2010" ;
– **(6) 95,74 %** : *Qui a gagné la ligue des champions en 2011 ?* "Barcelone en UEFA homme, Lyon en UEFA femme, Espérance de Tunis en CAF homme"
– **(7) 4,26 %** : *Quelle superbe victoire a remporté la France en 1998 ?* "1-0 contre la Finlande le 5 juin", "3-0 en France contre le Brésil le 12 juillet..." (onze victoires en tout en 1998) ;
– **(8) 10 %** : *Quand démarra le troisième gouvernement Fillon ?* "le 13/11/10" (annonce par Fillon), "le 16/11/10" (publication au Journal officiel) ;
– **(9) 4 %** : *Quel jour Nicolas Sarkozy est-il devenu président de la République ?* "Élu le 6 mai 2007, investi le 16 mai" ;
– **(10) 1 %** : *Quand fut fêté le bicentenaire de la révolution française ?* $1789 + 200 = 1989$;
– **(11) 11 %** : *Quels JO se sont déroulés il y a 16 ans ?* ;

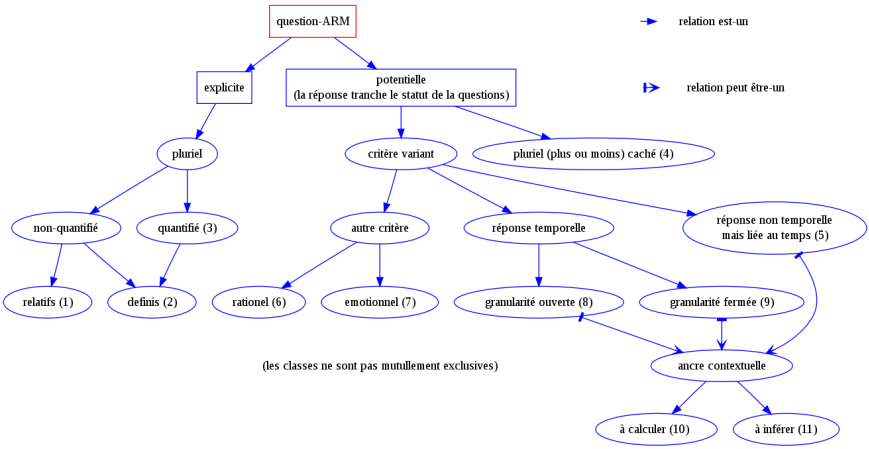

FIG. 1 – Typologie des questions-ARM. Les chiffres correspondent aux exemples précédents.

5.2 Approche classique avec FIDJI

Le SQR FIDJI permet de recouper les informations entre documents en se basant sur la syntaxe (Moriceau et Tannier, 2010) mais n'a pas encore de dispositif fonctionnel concernant les questions-ARM. Un traitement des question-liste existe cependant en recherchant dans un même document un groupe d'éléments consécutifs. Les résultats actuels vont nous servir de premier état des lieux pour implémenter le traitement des questions-ARM. Ainsi nous avons rencontré les phénomènes suivants :

- **(A)** FIDJI choisit de ne renvoyer qu'une seule réponse à fort score de confiance plutôt que plusieurs à faibles scores (même si la question est une question-liste) ;
- **(B)** FIDJI détecte la réponse-liste dans un document mais ne l'extrait pas correctement ;
- **(C)** FIDJI renvoie deux réponses-listes correctes sans les fusionner ;
- **(D)** FIDJI renvoie une réponse correcte ("2005") mais il existait une réponse correcte plus pertinente dans le passage ("octobre 2003") : **Quand est sorti l'Ibook G4 ?** *Avec les nouveaux iBook G4* ***2005****, Apple introduit Bluetooth2 de série (+ERD) (...). Le tableau ci-dessous retrace toute l'histoire de l'iBook G4 de sa sortie en* ***octobre 2003*** *à nos jours.*

Nous voyons donc des pistes concrètes d'améliorations puisque **(A)** est dû à un manque dans l'analyse des questions, **(B)** à une extraction à améliorer, **(C)** à un manque de recoupement entre les documents et **(D)** à une granularité de la pertinence de la réponse à renvoyer.

Le phénomène du critère variant est bien présent dans les résultats et nous montre l'intérêt à dépasser le cadre de la réponse unique à extraire d'un passage-candidat.

6 Conclusion et perspectives

En nous intéressant aux modes d'évaluation des SQR lors des campagnes pour le français, nous avons constaté un bridage nécessaire sur la présentation finale des réponses et relativement peu d'inter-documentalité pour les questions-listes. Après avoir constitué une collection de questions-ARM et de documents permettant d'y répondre, l'expérimentation avec un SQR rodé a confirmé la nécessité de mettre en place un traitement inter-document pour être en mesure de répondre le plus pertinemment possible à une question-ARM.

Nous allons donc implémenter un module de traitement des questions-ARM afin de dépasser le cadre habituel d'évaluation des SQR et se diriger vers un cadre utilisateur. En élargissant nos sources d'informations (HTML, éléments structuraux comme les tableaux), nous espérons bénéficier de plus d'informations pertinentes réparties dans des documents différents.

Un autre aspect intéressant est la présentation des réponses à l'utilisateur. Nous pensons proposer à l'utilisateur des réponses regroupées selon des critères variants s'ils existent, notamment à l'aide d'éléments structuraux (tableau par exemple). De plus, il serait intéressant d'ajouter aux réponses textuelles des données multimedia (URLs, images, etc.) qui permettront de justifier les réponses. L'évaluation des choix de regroupement serait alors faite du point de vue de l'utilisateur.

Plusieus approches applicatives se sont intéressées à la présentation des réponses à l'utilisateur. Par exemple, le SQR WolframQA[7] utilise également les images, les tableaux et les chronologies pour présenter plusieurs réponses à l'utilisateur. On retrouve les tableaux dans *Google squared*

[7] `http://www.wolframalpha.com`

(Crow, 2010) et des chronologies dans *Google News* et ChronoZoom[8] ainsi que dans les travaux de (Llorens *et al.*, 2011) qui s'intéresse à l'annotation temporelle de textes à des fins de visualisations ergonomiques pour l'utilisateur.

Références

AYACHE, C., GRAU, B. et VILNAT, A. (2006). Equer : the french evaluation campaign of question-answering systems. *In Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*.

AÏT-MOKHTAR, S., LUX, V. et BANIK, E. (2003). Linguistic parsing of lists in structured documents. *In Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003), Budapest, Hungary*.

BOS, J., GUZZETTI, E. et CURRAN, J. R. (2007a). The pronto qa system at trec 2007 : Harvesting hyponyms, using nominalisation patterns, and computing answer cardinality. *In TREC-16*.

BOS, J., GUZZETTI, E. et CURRAN, J. R. (2007b). The pronto qa system at trec 2007 : Harvesting hyponyms, using nominalisation patterns, and computing answer cardinality. *In TREC-16*.

BRAS, M., PRÉVOT, L. et VERGEZ-COURET, M. (2008). Quelle(s) relation(s) de discours pour les structures énumératives ? CMLF (Congrès mondial de linguistique française).

CHU-CARROLL, J., CZUBA, K., PRAGER, J. et BLAIR-GOLDENSOHN, S. (2004). Ibm's piquant ii in trec2004. *In TREC-13*.

CROW, D. (2010). Google squared : Web scale, open domain information extraction and presentation. *In ECIR*.

DAN I. MOLDOVAN AND, C. C. et BOWDEN, M. (2007). Lymba's poweranswer 4 in trec 2007. *In TREC-16*.

DANG, H. T., KELLY, D. et LIN, J. (2007). Overview of the trec 2007 question answering track. *In TREC-16*.

FANGTAO, L., XIAN, Z. et XIAOYAN, Z. (2008). Answer validation by information distance calculation. *In Coling 2008 : Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, IRQA '08, pages 42–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

FIGUEROA, A. et NEUMANN, G. (2008). Finding distinct answers in web snippets. *In In the 4th International Conference on Web Information Systems and Technologies*, pages 26–33. INSTICC Press.

GALA, N. (2003). *Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires*. Thèse de doctorat, Université Paris-Sud.

GATTERBAUER, W., BOHUNSKY, P., HERZOG, M., KRÜPL, B. et POLLAK, B. (2007). Towards domain-independent information extraction from web tables. *In Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 71–80. ACM.

GRAPPY, A. (2011). *Validation de réponse dans un système de question-réponse*. Thèse de doctorat.

HO-DAC, L.-M. (2007). *La position initiale dans l'organisation du discours : une exploration en corpus*. Thèse de doctorat, Université Toulouse le Mirail.

[8] `http://research.microsoft.com/en-us/projects/chronozoom/`

HO-DAC, L.-M., PÉRY-WOODLEY, M.-P. et TANGUY, L. (2010). Anatomie des structures énumératives.

JACQUEMIN, C. et BUSH, C. (2000). Fouille du web pour la collecte d'entités nommées. *In TALN*.

KAISSER, M. et BECKER, T. (2004). Question answering by searching large corpora with linguistic methods. *In TREC-13*.

KATZ, B. et LIN, J. (2003). Selectively using relations to improve precision in question answering. *In EACL-2003 workshop on natural language processing for question answering*.

KATZ, B., MARTON, G., FELSHIN, S., LORETO, D., LU, B., MORA, F., Özlem UZUNER, MCGRAW-HERDEG, M., CHEUNG, N., RADUL, A., SHEN, Y. K., LUO, Y. et ZACCAK, G. (2006). Question answering experiments and resources. *In TREC-15*.

LAIGNELET, M. (2009). *Analyse discursive pour le repérage automatique de segments obsolescents dans les documents encyclopédiques*. Thèse de doctorat.

LAIGNELET, M., KAMEL, M. et AUSSENAC-GILLES, N. (2011). Enrichir la notion de patron par la prise en compte de la structure textuelle - application à la construction d'ontologie. *In TALN*.

LLORENS, H., SAQUETE, E., NAVARRO, B. et GAIZAUSKAS, R. (2011). Time-surfer : time-based graphical access to document content. *In ECIR'11 : Proceedings of the 33rd European conference on Advances in information retrieval*, pages 767–771, Berlin, Heidelberg. Springer-Verlag.

LUC, C. (2001). Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. *In TALN*.

MOLDOVAN, D. I., CLARK, C. et BOWDEN, M. (2007). Lymba's poweranswer 4 in trec 2007. *In TREC-16*.

MORICEAU, V. et TANNIER, X. (2010). Fidji : Using syntax for validating answers in multiple documents. *Information Retrieval, Special Issue on Focused Information Retrieval*, (10791).

PÉRY-WOODLEY, M.-P. (2000). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. HDR.

QUINTARD, L., GALIBERT, O., ADDA, G., GRAU, B., LAURENT, D., MORICEAU, V., ROSSET, S., TANNIER, X. et VILNAT, A. (2010). Question answering on web data : the qa evaluation in quæro. *In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta*.

RAZMARA, M. et KOSSEIM, L. (2008). Answering list questions using co-occurrence and clustering. *In LREC*. European Language Resources Association.

SCHLAEFER, N., KO, J., BETTERIDGE, J., SAUTTER, G., PATHAK, M. et NYBERG, E. (2007). Semantic extensions of the ephyra qa system for trec 2007. *In TREC-16*.

TAJIMA, K. et OHNISHI, K. (2008). Browsing large html tables on small screens. *In UIST*, pages 259–268.

WANG, R. C., SCHLAEFER, N., COHEN, W. W. et NYBERG, E. (2008). Automatic set expansion for list question answering. *In EMNLP*.

WANG, Y. et HU, J. (2002). A machine learning based approach for table detection on the web. *In Proceedings of the 11th international conference on World Wide Web*, pages 242–250. ACM.

WU, M., ZHENG, X., DUAN, M., LIU, T. et STRZALKOWSKI, T. (2003). Questioning answering by pattern matching, web-proofing, semantic form proofing. *In TREC-12*, pages 578–585.