Évaluation de la réponse d'un système de question-réponse et de sa justification

Arnaud Grappy, Anne-Laure Ligozat, Brigitte Grau

LIMSI-CNRS
BP 133 ORSAY CEDEX
prenom.nom@limsi.fr

RÉSUMÉ. *Les systèmes de question-réponse fournissent une réponse à une question en l'extrayant d'un ensemble de documents. Avec celle-ci ils fournissent également un passage de texte permettant de la justifier. On peut alors chercher à évaluer si la réponse proposée par un système est correcte et justifiée par le passage. Pour cela, nous nous sommes fondés sur la vérification de différents critères : le premier tient compte de la proportion et du type des termes communs au passage et à la question, le second de la proximité de ces termes par rapport à la réponse, le troisième compare la réponse à considérer avec celle obtenue par le système de question-réponse* FRASQUES *utilisé sur le passage à juger et le dernier est une vérification du type de la réponse. Les différents critères sont ensuite combinés grâce à un classifieur utilisant les arbres de décision.*[1]

ABSTRACT. *Question answering systems extract precise answers from a set of documents, and return the answers along with text snippets which justify them. It is thus possible to assess the answer returned by a system, with respect to its snippet. In this article, such answer validations are processes, according to different criteria: the first one estimates the proportion of common terms in the snippet and the question, the second one measures the proximity of these terms and the answer, the third one compares the answer to judge with this returned by our own question answering system* FRASQUES*, and the last one checks the type of the answer. The different criteria are then combined thanks to a classifier using decision trees.*

MOTS-CLÉS : *Justification de réponses, Système de question-réponse, Recouvrement lexical*

KEYWORDS: *Answer justification, Question answering system, Lexical overlap*

1. travail effectué dans le cadre de CONIQUE, projet ANR-05-BLAN-0085-01

1. Introduction

Les systèmes de question-réponse (SQR) recherchent et extraient des textes la réponse à une question posée en langue naturelle, si celle-ci figure dans la collection interrogée. Les questions traitées sont de nature factuelle et appellent à une réponse concise, tenant en peu de mots. Les exemples typiques sont les questions dont la réponse est une entité nommée, comme par exemple une personne dans le cas de la question «Qui a provoqué la faillite de la banque Barings ?» dont la réponse est «Nick Leeson», ou bien une organisation, comme pour «Which US Army Division provided the paratroopers who took part in the invasion of Haiti ?», qui a pour réponse «the 82nd Airborne Division». Les textes sources sont issus de collections de journaux, et questions et textes sont correctement rédigés.

La méthodologie appliquée pour rechercher ces réponses consiste à appliquer des filtres successifs sur les textes jusqu'à ne retenir qu'une réponse : filtres sur la collection pour ne retenir qu'un sous-ensemble de documents, filtres sur la taille des textes retenus pour ne retenir que des passages de texte, filtres sur les éléments des phrases pour ne retenir que la réponse. Ces filtres sont formés à partir des résultats de l'analyse des questions et portent sur les termes à rechercher, les relations entre eux ainsi que le type de réponse cherché.

Quand les systèmes proposent une réponse, ils ont à évaluer si les extraits candidats issus de ces filtres sont corrects ou non, et pour cela se fondent sur une évaluation de la correspondance entre le passage candidat et la question ; si, à l'issue de l'évaluation, la réponse est considérée comme correcte, ce passage constitue une justification de la réponse. Cette évaluation peut être obtenue à l'issue d'un processus de déduction permettant de relier la réponse et le passage qui la contient aux informations données dans la question ; alors le processus lui-même garantit que la réponse trouvée répond bien à la question (Tatu *et al.,* 2006a). L'évaluation peut également se fonder sur la recherche d'une paraphrase de la question dans le passage, et des mesures de similarité entre les deux entités. Si l'on reprend la question «Which **US Army Division** provided the **paratroopers** who **took** part in the **invasion** of **Haiti** ?», notre système sur l'anglais propose les passages suivants :

– «A light-infantry contingent of 2,000 from the **Army** 's 10th Mountain **Division** is being sent later this week aboard the Eisenhower to the **Haiti** shoreline, where the *troops* will wait " within eyesight " of the beach and conduct mop-up and peacekeeping operations on the island nation should any **invasion take** place.»

– «Airplanes carrying **paratroopers** of the 82nd Airborne **Division** were already on their way to Port-au-Prince when **Haiti** 's military dictators finally accepted **U.S.** demands that they voluntarily step down or be ousted by force.»

Même si le premier passage contient plus de termes de la question, à l'identique ou sous forme approchée comme pour « troop », le second contient le terme exact de la question qui joue un rôle prépondérant. Ce second passage constitue bien une justification de la réponse, mais fait appel à des connaissances externes au passage.

Les connaissances nécessaires concernent souvent les reformulations de termes par des synonymes. D'autres connaissances lexicales permettent de lier les termes selon des relations de causalité ou d'en obtenir une définition. Enfin, le dernier type de connaissance fait appel à des connaissances générales sur le monde.

Le passage justificatif est en général présenté à l'utilisateur afin qu'il puisse vérifier que la réponse trouvée par le système est correcte. Il est à noter, que selon l'utilisateur, et ses connaissances, il sera considéré ou non comme une justification.

Dans cet article, nous avons cherché à évaluer le degré de justification d'un passage, par rapport à une réponse proposée, indépendamment du système qui la propose. Les critères que nous prenons en compte sont ceux que notre système de question-réponse sait reconnaître automatiquement, critères que nous avons combinés à l'aide d'un classifieur. Les résultats ont été évalués sur les données de la campagne AVE 2006[1].

2. État de l'art

Les méthodes de validation de réponse par évaluation du passage justificatif sont présentes aux évaluations RTE (Recognising Textual Entailement) et AVE. La tâche de RTE consiste à dire si un passage prouve une hypothèse alors que la deuxième consiste à dire si un passage justifie la réponse donnée à une question. Cette dernière tâche peut être ramenée à la première en considérant le couple question + réponse comme une hypothèse. Les méthodes employées sont de trois natures :

– méthodes intégrant différents critères : recouvrement lexical, équivalences syntagmatiques et paraphrase locale ;

– méthodes d'appariement de graphes ;

– méthodes de preuve logique.

Les méthodes intégrant différents critères reposent sur un classifieur, en général un arbre de décision. Les critères varient selon les systèmes mais correspondent aux trois types cités ci-dessus. Le recouvrement lexical (Newman *et al.,* 2005), (Hickl *et al.,* 2006b) mesure la plus longue sous-chaîne commune entre l'hypothèse et le passage. Ce critère consistant à favoriser les passages les plus compacts qui contiennent les termes de la question a toujours fait partie des critères de sélection de passage dans les SQR.

Les autres critères consistent à mettre en correspondance les termes, en tenant compte des variations morphologiques et sémantiques (Herrera *et al.,* 2006), en allant jusqu'à la recherche de paraphrases locales et la vérification de relations syntagmatiques entre prédicat et argument (Hickl *et al.,* 2006a).

Les méthodes d'appariement de graphes visent à mettre en superposition la représentation de l'hypothèse (ou de la question) et celle du passage. Elles opèrent

1. Answer Validation Exercice at CLEF : http ://www.clef-campaign.org

soit au niveau syntaxique (Kouylekov *et al.,* 2006), soit au niveau sémantique (Glöckner, 2007). Elles requièrent de savoir construire une représentation complète des phrases, capacité nécessairement mise en échec quand il s'agit de produire la représentation sémantique de toute phrase, et réduite quand il s'agit d'analyse syntaxique, car les analyseurs sont souvent mis en échec lorsqu'il s'agit de rattacher les groupes. (de Salvo Braz *et al.,* 2005) mêlent ces deux approches en travaillant sur différentes représentations.

Les méthodes par preuve logique (Tatu *et al.,* 2006b), si elles fournissent des résultats sûrs, sont en général peu robustes, puisqu'elles dépendent de la complétude de bases de connaissances. De ce fait, (Tatu *et al.,* 2006b) ajoute l'utilisation d'une méthode d'alignement lexical et améliore ainsi très nettement ses résultats.

En conclusion, les meilleurs systèmes sont ceux qui intègrent différents critères et différentes méthodes. Pour notre part, nous avons voulu mesurer l'intégration de critères lexico-sémantiques et de relations syntagmatiques obtenues grâce à l'application de notre SQR sur les passages. Nous avons de ce fait étudié plus en détail l'importance des termes de la questions, selon leur catégorie morpho-syntaxique, mais surtout selon le rôle sémantique qu'ils jouent dans la formulation de la réponse.

3. Le système de validation

Dans cet article, nous avons donc cherché à évaluer des réponses à des questions, en fonction du passage justificatif associé à chaque réponse. La validation des réponses repose sur les critères reconnus par notre système de question-réponse FRASQUES pour le français : ce sont donc les différents critères utilisés par ce système pour déterminer la validité d'une réponse qui sont évalués. L'architecture du système de validation des réponses est présentée figure 1.

Trois modules du système de question-réponse sont utilisés pour analyser les passages justificatifs : l'analyse de la question, le traitement des documents et l'extraction de la réponse.

Dans un premier temps, la question est analysée, afin de déterminer un certain nombre de ses caractéristiques, comme la catégorie de la question (par exemple, la question « Qu'est-ce que l'UNITA ?» est de la catégorie « définition »), ou le type attendu de la réponse, qui peut être une entité nommée (« LOCATION-CITY » pour « Dans quelle ville Mozart était-il né ? ») ou un type général (« traité » pour « Quel traité a été signé en 1979 ? »).

Puis, les passages justificatifs sont analysés : les variantes des mots de la question sont recherchées par l'outil FASTR (Jacquemin, 1996), qui permet de reconnaître des variations morphologiques, sémantiques ou syntaxiques, de termes ; les entités nommées sont également étiquetées.

Enfin, cet étiquetage en entités nommées et des patrons syntaxiques permettent d'extraire des réponses candidates de ces passages.

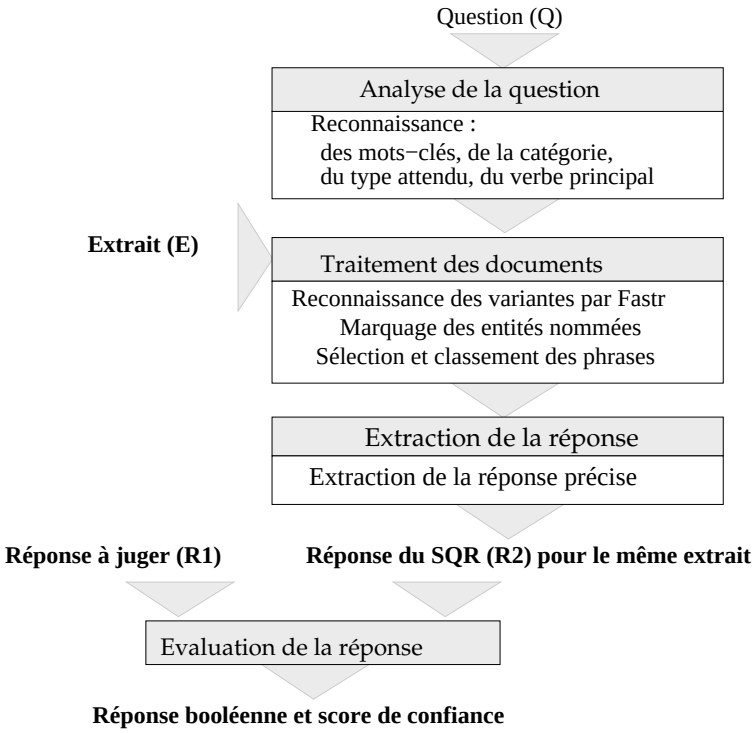

Figure 1. *Système de validation de réponses*

Le système de validation va ensuite pouvoir évaluer la réponse à juger (R1) avec sa justification (Extrait E), en fonction des différents critères issus de cette chaîne de traitements : termes communs à la question et à la justification ou comparaison de la réponse trouvée par FRASQUES (R2) et celle à juger (R1) par exemple (Ligozat *et al.,* 2007). Il retourne donc OUI si la réponse est considérée comme correcte et justifiée par le passage E, et NON sinon.

Notre objectif dans cet article est d'évaluer l'importance de chacun de ces critères, afin de déterminer les connaissances et les processus nécessaires à mettre en oeuvre pour extraire ou évaluer la justification d'une réponse. Les critères évalués ici sont de différentes natures : comparaison des mots communs à la question et à la justification de la réponse, comparaison de la réponse à évaluer avec celle fournie par notre SQR FRASQUES, calcul de la plus longue chaîne de mots commune à la question et à la justification, et vérification du type de la réponse grâce à Wikipédia. Pour effectuer cette étude, nous nous sommes appuyés sur un corpus issu de la campagne d'évaluation AVE présentée ci-dessous, ce qui nous a permis d'étudier ces critères indépendamment de notre stratégie de recherche de la réponse.

4. Le corpus d'évaluation

La campagne AVE organisée pour la première fois en 2006 a permis la création d'un corpus d'évaluation constitué d'un ensemble de paires (hypothèse, passage), où l'hypothèse réunit la question et la réponse. Cette campagne était proposée dans sept langues différentes telles que l'anglais, l'espagnol ou le français et a rassemblé treize participants. Le corpus sur lequel porte l'étude est en français. Celui-ci se compose de 190 questions et de 3267 couples (réponses à ces questions, passage justificatif). Les réponses ont été évaluées par les organisateurs : 2358 ont été jugées incorrectes (NON) 627 seulement correctes (OUI), et 280 n'ont pas été jugées (INCONNU). Les réponses non jugées ne seront pas considérées car elles ne permettent pas d'évaluer nos résultats sur le même référentiel que celui de l'évaluation.

4.1. *Mesures d'évaluation*

Pour chacune des paires, les systèmes doivent déterminer si la réponse est justifiée par le passage en retournant la valeur OUI et la valeur NON sinon, avec un score de fiabilité. Pour évaluer les systèmes, seules les réponses positives sont comptabilisées en utilisant trois critères : la précision, le rappel et la f-mesure :

– la précision est la proportion de réponses correctes parmi les OUI donnés par le système ;

– le rappel est le rapport entre le nombre de réponses OUI correctes retournées par le système et le nombre de réponses OUI attendues ;

– la f-mesure permet de combiner ces deux paramètres.

$$précision = \frac{\#\ réponses\ \text{OUI}\ correctes\ renvoyées}{\#\ réponses\ \text{OUI}\ données}$$

$$rappel = \frac{\#réponses\ \text{OUI}\ correctes\ renvoyées}{\#\ réponses\ \text{OUI}\ attendues}$$

$$f - mesure = \frac{2 * précision * rappel}{précision + rappel}$$

4.2. *Présentation du corpus*

Deux cas simples permettent de mieux appréhender la base et la répartition des réponses correctes justifiées par rapport à l'ensemble des réponses fausses. Le premier consiste à toujours répondre OUI et le second consiste à répondre OUI à une réponse sur deux de manière aléatoire. Le tableau 1 présente les résultats obtenus à ces tests.

méthode	précision	rappel	f-mesure
toujours OUI	0.22	1	0.36
50 % de OUI	0.23	0.5	0.31

Tableau 1. *Tests de base*

4.3. *Suppression des réponses* NON *sûres*

Parmi les réponses négatives, certaines le sont pour des raisons évidentes. Voici l'ensemble de ces raisons :

– le passage ne contient pas la réponse ;

– la réponse est contenue dans la question ;

– le passage est contenu dans la réponse ;

– la date de la réponse ne peut pas être trouvée dans le passage et ne correspond pas non plus à sa date de création ;

– l'entité attendue en réponse ne correspond pas à celle obtenue ; ce cas se rencontre par exemple quand la question attend une personne comme réponse et qu'une date est obtenue.

Les traitements mis au point pour supprimer des propositions fausses ont permis de détecter 995 cas de réponses invalides. Parmi elles, il y a en réalité 965 réponses négatives pour 30 réponses correctes. Nous avons donc considéré que l'on pouvait travailler sur un corpus restreint, dans lequel ces réponses NON évidentes ont été supprimées. Par la suite les calculs seront effectués sur ce corpus restreint. Le tableau 2 montre la répartition des résultats avant et après élimination des réponses NON évidentes.

corpus	nombre de OUI	nombre de NON	nombre total de réponses
avant suppression	627	2358	2985
après suppression	597	1393	1990

Tableau 2. *Composition du corpus restreint*

5. Méthodologie utilisée

Afin d'évaluer la justification, nous avons utilisé une approche par apprentissage. Un ensemble de critères est fourni à un classifieur qui en cherche une combinaison permettant de décider si le passage justifie ou non la réponse. Une partie de la base de départ constitue la base d'apprentissage et l'autre a été gardée afin d'effectuer les tests. Elles correspondent à la répartition suivante : $\frac{3}{4}$ des données servent à la base d'apprentissage et $\frac{1}{4}$ à la base de test. La base de test est donc composée de 578 réponses parmi lesquelles 162 sont valides et 416 ne le sont pas.

La base d'apprentissage contenant beaucoup plus de réponses invalides que de réponses valides, afin de maximiser les résultats, le nombre de réponses invalides a été légèrement diminué.

L'autre choix à faire est celui du classifieur. Le programme WEKA[2] permet d'utiliser un grand nombre de classifieurs allant des SVMs aux réseaux de neurones en pas-

2. WEKA : http ://www.cs.waikato.ac.nz/ml/weka

sant par les systèmes bayésiens. D'un point de vue pratique, la combinaison d'arbres de décision grâce à la méthode bagging a donné les meilleurs résultats. Dans cette méthode, les résultats obtenus par les différents arbres sont réunis par vote afin de proposer un seul résultat.

Les sections suivantes exposent la prise en compte des quatre critères suivants, et les évaluent de manière individuelle avant de chercher à les combiner :

– la correspondance de termes communs au passage et à la question ;

– la correspondance de la réponse et de celle fournie par un système de question-réponse recherchant la réponse dans le passage ;

– la proximité de ces termes dans le passage ;

– la vérification du type de la réponse grâce à Wikipédia.

6. Étude des termes

6.1. *Proportion de termes communs*

Une première étude consiste à considérer le taux de mots de la question présents dans le passage. Si le taux est supérieur à un seuil la réponse sera considérée comme valide. Ce critère est fondé sur l'hypothèse suivante : un passage justifie la réponse à une question s'il possède un certain nombre de termes communs ou similaires avec elle, dans la mesure où cela signifie que le passage a de fortes chances de parler du thème de la question, et donc de posséder la réponse cherchée.

Les mots du passage sont reconnus par FASTR ce qui permet de reconnaître les variantes ou les synonymes. Notons que certaines catégories de mots non pertinentes ne sont pas considérés. Ces catégories sont les déterminants, les prépositions et les adverbes. Les résultats ainsi obtenus sont présentés dans le tableau 3.

précision	rappel	f-mesure
0.50	0.71	0.59

Tableau 3. *Résultats de la validation des réponses, en considérant les termes communs au passage et à la question*

Ces résultats constituent en fait un test de référence, puisque ce critère est présent dans tous les systèmes.

6.2. *Importance des termes*

6.2.1. *Catégories morpho-syntaxiques*

Parmi les mots communs au passage et à la question, une hypothèse peut être faite sur leur importance en fonction de leur catégorie morpho-syntaxique : la présence d'un nom propre commun est sans doute plus importante que celle d'un adjectif.

Nous avons donc séparé les mots communs en fonction de leurs catégories morpho-syntaxiques. Seuls les noms propres, les noms communs, les adjectifs, les verbes et les nombres sont considérés dans cette partie.

Une valeur de pertinence est donnée à chaque catégorie conservée. Cette valeur correspond à la proportion de mots de la catégorie de la question présents dans le passage. Si la question ne contient pas de mot de cette catégorie la valeur sera de -1. Le tableau 4 montre l'effet de chaque critère sur la réponse.

critère	précision	rappel	f-mesure
nom communs	0.34	0.86	0.49
noms propres	0.33	0.88	0.48
nombres	0.31	0.85	0.45
verbes	0.29	0.75	0.41
adjectifs	0.52	0.23	0.32

Tableau 4. *Évaluation des catégories morpho-syntaxiques*

Nous pouvons constater que les critères autres que adjectif obtiennent un fort rappel et une faible précision ce qui indique que les réponses correctes sont souvent reconnues mais aussi que de nombreuses réponses sont considérées comme correctes à tort. Par ailleurs il n'y a pas une catégorie qui ressort par rapport aux autres.

Après avoir obtenu ces résultats, nous avons regardé l'intérêt de ces critères combinés avec la proportion de mots communs. Ils n'ont pas permis d'améliorer significativement les résultats.

6.2.2. *Termes selon leur rôle dans la question*

D'autres critères plus sémantiques portant sur les termes peuvent être considérés. Ces critères prennent en compte le rôle d'un mot dans la question et sont les suivants :

– **le focus :** l'entité sur laquelle porte la question et pour laquelle une caractéristique ou une définition est demandée. Par définition, cet élément devrait être repris dans le passage pour exprimer la réponse. Dans la question « Quel est le sport pratiqué par Zinédine Zidane ?», le focus est « Zinédine Zidane » ;

– **le type attendu :** n'est pas toujours présent, mais vient souvent préciser le type de la réponse quand il l'est. Dans la question précédente le type est « sport » ;

– **le verbe principal :** c'est le verbe présent dans la question et ayant un rôle important dans la formulation de la réponse quand il introduit un fait, une action. Il

n'est donc ni un auxiliaire, ni un verbe modal et ne sert pas non plus à poser la question de manière indirecte. Dans la question précédente le verbe principal est « pratiquer » ;

– **les bitermes :** un biterme est une suite de deux mots reconnus comme étant liés syntaxiquement, comme « prix Nobel »ou« premier prix ». Trouver un bi-terme de la question dans le passage signifie souvent que les mots présents sont utilisés dans le même sens. Nous pouvons nous limiter à deux mots, car nous extrayons tous les bi-termes des groupes nominaux de la question et un terme formé de trois mots par exemple sera décomposé en deux bi-termes.

Ces critères ont été évalués en étudiant leur présence commune dans le passage et la question. Le tableau 5 montre les résultats obtenus pour chacun.

critère	précision	rappel	f-mesure
focus	0.5	0.7	0.59
type attendu	0.29	0.66	0.4
verbe principal	0.32	0.54	0.41
biterme	0.42	0.42	0.42

Tableau 5. *Évaluation des mots importants de la question*

Nous pouvons remarquer que le focus est un critère plus intéressant que les autres. Le type de réponse attendu obtient un bon rappel, mais a une précision très faible, ce qui ne le rend pas très discriminant. L'étape suivante a consisté à combiner l'ensemble des critères vus dans cette section afin de voir si ils sont plus pertinents ensemble. Le tableau 6 donne les résultats obtenus.

précision	rappel	f-mesure
0.5	0.8	0.62

Tableau 6. *Validation des réponses avec une combinaison des critères*

Ainsi, l'utilisation des rôles sémantiques de la question permet d'améliorer le rappel.

7. Vérification de la réponse

Le second type de critère consiste à tenir compte de la réponse trouvée par notre SQR dans le passage. Si la réponse renvoyée est la même que celle à juger, elle a de bonnes chances d'être correcte.

Dans notre SQR, la méthode d'extraction de la réponse dépend du type de question. Si la question attend en réponse une entité nommée, le système extraira l'entité du bon type la plus proche des mots de la question. Sinon, la réponse est recherchée grâce à des patrons d'extraction. Ces patrons sont articulés autour du focus, du verbe principal ou du type général et correspondent à des règles syntaxiques locales. Ils

permettent de vérifier l'existence de la relation attendue entre la réponse et l'élément pivot retenu, quand il s'agit du focus ou du verbe, ou de vérifier le type de la réponse quand le pivot est le type.

La réponse obtenue est comparée à la réponse à tester en comptabilisant le nombre de mots communs aux deux réponses et en divisant ce résultat par le nombre de mots des réponses. Cette méthode obtient les résultats présentés dans le tableau 7.

précision	rappel	f-mesure
0.4	0.7	0.5

Tableau 7. *Comparaison des réponses*

8. Plus Longue Chaîne Commune (LCC)

La méthode présentée ici calcule la plus grande chaîne de mots consécutifs de l'hypothèse se trouvant dans le passage.

Cette notion permet d'approximer la vérification de relations syntagmatiques : la plus longue chaîne commune correspond à une formulation partielle de l'information analogue à celle de la question. L'hypothèse formée pour rechercher cette chaîne correspond à la forme affirmative de la question à laquelle la réponse est ajoutée. L'exemple suivant sera utilisé dans la suite de notre explication :

> **Question** : Qui est le père de la reine Elisabeth 2 ?
> **Passage** : **Georges VI**, **le père** d' **Elisabeth 2**, l'actuelle **reine** d'Angleterre ...
> **Réponse** : Georges VI
> ***Hypothèse*** : *Georges VI est le père de la reine Elisabeth 2*.

8.1. *Algorithme*

L'algorithme recherche la plus longue chaîne de mots consécutifs mais non ordonnés présents dans le passage et l'hypothèse et ce quelque soit la question.

Tout d'abord, afin de faciliter le rapprochement entre l'hypothèse et le passage, les mots sont normalisés : les variantes reconnues sont ramenées au terme d'origine. L'algorithme identifie ensuite tous les plus grands groupes de mots adjacents communs à la question et à l'hypothèse, appelés aussi éléments. Dans l'exemple, les groupes « Georges VI », « le père », « Elisabeth 2 » et « reine » sont ainsi reconnus.

Ensuite, une chaîne est initialisée à partir de chacun des groupes obtenus puis agrandie récursivement par un élément si celui-ci est soit directement adjacent soit séparé par un certain nombre d'items autorisés, comme une virgule, un déterminant, et un seul éventuel mot non vide considéré comme bonus. A la fin, seule la plus grande

chaîne est conservée. Dans l'expression « Elisabeth 2, l'actuelle reine », « Elisabeth 2 » et « reine » sont reliés car ils ne sont séparés que par un mot non vide (« actuelle »). La chaîne obtenue pour l'exemple est donc : « Georges VI, le père de Elisabeth 2 reine ».

La formule suivante explique la chaîne calculée de manière plus formelle :
chaîne $= m_1...m_n | \forall m_i, m_i \in$ *hypothèse et* $m_i n^* l n^* m_{i+1} \in passage$ *avec n* un ensemble de mots vides et l un éventuel mot non vide.*

Pour avoir une idée de la validité de la réponse, un poids est associé à la chaîne obtenue. Ce poids correspond au rapport entre le nombre de mots de la chaîne calculée et le nombre de mots de l'hypothèse. Le poids associé à l'exemple est de 0.8, car la chaîne calculée contient 8 mots, tandis que l'hypothèse en contenait 10.

Ce critère a aussi été introduit dans d'autres systèmes. Dans (Bosma *et al.,* 2006), l'hypothèse est modifiée afin d'être la plus proche possible du passage. Pour ce faire, le système remplace les paraphrases de l'hypothèse présentes dans le passage par leurs expressions associées. Puis il recherche la plus longue chaîne de mots non forcément consécutifs présents dans le passage et dans la nouvelle hypothèse et, pour détecter l'implication, compare la taille en nombre de mots de cette chaîne à celle de l'hypothèse.

On le retrouve aussi dans les SQR, par exemple (Laurent Gillard, 2006), qui calcule la distance, en nombre de mots, entre les mots de la question et la réponse. Cette mesure ne s'applique que si la réponse attendue est une entité nommée.

Lorsque l'on applique ce critère pour déterminer si le passage justifie ou non l'hypothèse, on trouve une frontière de décision aux alentours de 0.54. Il faut donc qu'au moins la moitié des mots de l'hypothèse se trouvent dans le passage assez proches les uns des autres pour que la réponse soit validée. Ce seuil entraîne les résultats présentés dans le tableau 8.

précision	rappel	f-mesure
0.53	0.80	0.64

Tableau 8. *Plus longue chaîne commune*

9. Vérification du type de la réponse grâce à Wikipédia

Pour un certain nombre de questions, la réponse attendue est d'un type particulier. Jusqu'à présent, la vérification de ce type s'effectuait en étudiant sa présence dans le passage candidat et s'il était présent la réponse était considérée comme étant du bon type. Toutefois cette vérification est sujette à deux types d'erreurs :

– si le type n'est pas contenu dans le passage, la réponse peut néanmoins être correcte. La relation de type est alors implicite ;

– si le passage contient effectivement le type de la question, rien n'implique que la réponse soit correcte, le type ne qualifiant pas nécessairement la réponse.

La méthode présentée ici a pour but de pallier ces problèmes, en s'appuyant sur l'encyclopédie en ligne Wikipédia[3]. L'idée est la suivante : si le type général est trouvé dans la page Wikipédia associée à la réponse, cette dernière a de fortes chances d'être une instance de ce type.

La méthode consiste à tester la présence du type, pris sous sa forme textuelle, dans les pages Wikipédia ayant comme titre la réponse ou dont le titre contient la réponse. Si celle-ci est trouvée, le poids faisant office de critère est de 1. Dans le cas contraire, la réponse est supposée ne pas être du type considéré et le poids sera de 0. Notons que pour les questions ne précisant pas le type de la réponse, aucune recherche n'est à effectuer et le poids sera de -1.

9.1. *Test*

Afin de tester ces résultats, une première étude portant sur la proportion de réponses correctes fournies par Wikipédia a été effectuée et les résultats suivants ont été obtenus : quand la réponse est du type donné (comme « Albert Einstein » pour le type « physicien ») le système le reconnaît également à 94 %. Quand la réponse n'est pas de ce type, (« Roosevelt » et « physicien ») le système le détecte à 88 %.

Le test suivant a consisté à étudier l'effet de ce critère sur la validité de la réponse. Le tableau 9 montre la répartition des réponses pour les questions typées dans les cas où le type général est trouvé et dans les cas où il ne l'est pas.

	réponse correcte	réponse fausse
type trouvé	35 %	65 %
type non trouvé	17 %	83 %

Tableau 9. *Vérification du type de la réponse avec Wikipédia*

Nous pouvons constater que quand le type est trouvé, il est difficile de faire une estimation fiable, mais si ce n'est pas le cas, la réponse a de bonnes chances d'être fausse.

10. Combinaison des critères

Nous avons vu dans les sections précédentes qu'un certain nombre de critères jouent un rôle dans l'évaluation des réponses et leur justification ; rappelons les :

– la proportion de termes communs au passage et à la question ;

3. Wikipédia : http ://fr.wikipedia.org

– la proportion de noms propres, de noms communs, de nombres et d'adjectifs communs au passage et à la question ;

– la présence commune au passage et à la question du focus, du type de la question, de bitermes et du verbe principal ;

– la correspondance de la réponse à juger avec la réponse obtenue par notre système de question-réponse passé sur le passage à considérer ;

– la taille de la plus grande chaîne commune au passage et à l'hypothèse ;

– la vérification du type grâce à Wikipédia.

Le tableau 10 rappelle les résultats obtenus par chacun des critères.

critère	précision	rappel	f-mesure
Proportion de termes communs	0.50	0.71	0.59
Apprentissage sur termes	0.5	0.8	0.62
Vérification de la réponse	0.4	0.7	0.5
LCC	0.53	0.8	0.64

Tableau 10. *Ensemble des critères*

Nous avons vu que si ces critères sont pertinents, ils ne permettent pas, pris seuls, d'emporter la décision. Aussi, ces critères ont été rassemblés en utilisant une combinaison d'arbre de décision grâce à la méthode bagging, ce qui a permis d'obtenir les résultats présentés dans le tableau 11.

précision	rappel	f-mesure
0.58	0.82	0.68

Tableau 11. *Combinaison des critères*

L'utilisation d'un arbre de décision permet de voir l'apport respectif de chacun des critères. Ici le critère prédominant est la taille de la plus grande chaîne commune puis viennent la correspondance de la réponse et le nombre de termes communs. Au troisième niveau on trouve la vérification du type grâce à Wikipédia et le focus. L'utilisation de ce classifieur a montré que la présence du verbe principal n'était pas un indice pertinent.

En utilisant une validation croisée, les résultats présents dans le tableau 12 sont obtenus. Cette validation consiste à découper les données en 10 groupes puis à effectuer un apprentissage sur les 9 premiers, à tester sur le dernier. Cette étape est répétée 10 fois puis le résultat final est obtenu en moyennant les différents résultats.

précision	rappel	f-mesure
0.66	0.73	0.69

Tableau 12. *Validation croisée*

Nous pouvons constater que les résultats globaux sont analogues à ceux obtenus par la validation précédente, avec un rappel plus faible compensé par une meilleure précision .

Afin d'évaluer les résultats obtenus, il est possible de les comparer avec ceux des autres systèmes candidats à AVE. Précisons que notre travail est postérieur à l'évaluation et s'est effectué à partir d'une partie de son corpus. Le tableau suivant permet de comparer le système présenté jusqu'ici avec celui obtenant les meilleurs résultats sur le français (MLENT) (Kozareva *et al.,* 2006) et celui obtenant les meilleurs résultats toutes langues confondues (COGEX) (Tatu *et al.,* 2006a).

La méthode MLENT est une combinaison d'une approche par apprentissage et d'une approche par logique. L'approche par apprentissage a comme critères la proportion de termes communs, la plus grande chaîne de mots consécutifs présents dans le passage et l'hypothèse et la proportion de mot différents du passage à l'hypothèse en suivant le même ordre. La méthode COGEX suit une approche logique.

système	précision	rappel	f-mesure
MLENT	0.34	0.73	0.57
COGEX	0.53	0.78	0.63
notre système	0.66	0.73	0.69

Tableau 13. *Comparaison des résultats avec ceux des autres systèmes*

Nous voyons donc que les résultats obtenus par le système présenté ici sont comparables à ceux des autres systèmes.

11. Conclusion

Le système réalisé obtient de très bons résultats en considérant essentiellement des critères lexicaux et des paraphrases locales. Parmi les critères que nous avons étudiés, la prise en compte des rôles de certains termes de la question n'avait pas été testée dans d'autres systèmes. Or, on peut constater un apport réel dû à ces rôles. Nous avons ainsi un système qui se comporte bien, et dans lequel nous pouvons rajouter d'autres critères.

Il nous reste à approfondir le rapprochement entre hypothèse et question lorsqu'il y a présence d'inférences causales entre termes, ou lorsqu'un terme est remplacé par sa définition.

Par ailleurs, il faudra aussi approfondir la vérification des relations entre termes, et ne pas s'en tenir à la notion de chaîne commune uniquement.

Ce travail a aussi pour but d'améliorer notre système de question-réponse et l'évaluation de la validité de la réponse par rapport au passage justificatif va pouvoir être ajoutée et testée dans le SQR et son apport mesuré par rapport aux choix existants.

12. Bibliographie

Bosma W., Callison-Bursh C., « Paraphrase substitution for Recognizing textual entailment », *Working Notes for the CLEF 2006 Workshop (AVE)*, 2006.

de Salvo Braz R., Girju R., Punyakanok V., Roth D., Sammons M., « An Inference Model for Semantic Entailment in Natural Language », *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

Glöckner I., « University of Hagen at CLEF 2007 : Answer Validation Exercise », *Working Notes for the CLEF 2007 Workshop (AVE)*, 2007.

Herrera J., Rodrigo A., Penas A., Verdejo F., « UNED submission to AVE 2006 », *Working Notes for the CLEF 2006 Workshop (AVE)*, 2006.

Hickl A., Williams J., Bensley J., Kirk Roberts Y. S., Rink. B., « Question Answering with LCC's Chaucer at TREC 2006 », *Proceedings of The Fifteenth Text Retrieval Conference (TREC 2006)*, 2006a.

Hickl A., Williams J., Bensley J., Roberts K., Rink B., Shi Y., « Recognizing Textual Entailment with LCC's GROUNDHOG System », *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006b.

Jacquemin C., *A symbolic and surgical acquisition of terms through variation*, Springer, Heidelberg, p. 425-438, 1996.

Kouylekov M., Negri M., Magnini B., Coppola B., « Towards Entailment-based Question Answering : ITC-irst at CLEF 2006 », *Working Notes for the CLEF 2007 Workshop (AVE)*, 2006.

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 Gillard Patrice Bellot M. E.-B., *Influence de mesures de densité pour la recherche de passages et l extraction de réponses dans un système de questions-réponses*. 2006.

Ligozat A.-L., Grau B., Robba I., Vilnat A., « Systèmes de questions-réponses : vers la validation automatique des réponses », *Actes de la 14è Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, TALN, Toulouse, 2007.

Newman E., Stokes N., Dunnion J., Carthy J., « UCD IIRG Approach to the Textual Entailment Challenge », *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, p. 53-56, 2005.

Tatu M., Iles B., Moldovan D., « Automatic Answer Validation Using COGEX », *Working Notes for the CLEF 2006 Workshop*, Springer, 2006a.

Tatu M., Iles B., Slavick J., Novischi A., Moldovan D., « COGEX at the Second Recognizing Textual Entailment Challenge », *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006b.