Intérêts d'un Corpus issu du Web pour les Systèmes Question-Réponse

Véronique Moriceau, Farida Aouladomar
Institut de Recherches en Informatique de Toulouse
118, route de Narbonne, 31062 Toulouse cedex 9
{moriceau, aouladom}@irit.fr

Aujourd'hui, le Web met à la disposition du grand public un très grand nombre de données et les systèmes de recherche d'informations sont des outils à première vue pratiques pour qui souhaite trouver une réponse à une question. En revanche, là où les dictionnaires et encyclopédies fournissent une réponse unique, synthétique et cohérente, les moteurs de recherche actuels proposent à l'utilisateur un ensemble de liens vers des pages Web et/ou des extraits de ces pages traitant du thème de la requête. Il revient à l'utilisateur d'en sélectionner les plus intéressants et de rechercher au sein des textes la réponse à sa question. Le temps que nécessite cette démarche laborieuse est souvent long. De plus, ces systèmes ne cherchent pas à connaître le sens de la question, ils renvoient parfois à des documents qui n'ont pas de rapport direct avec celle-ci.

Les systèmes question-réponse coopératifs proposent une alternative à ce problème. En effet, ces systèmes recherchent un ensemble de pages traitant de la question posée et proposent à l'utilisateur une réponse unique, celle que le système juge la "meilleure". Nous nous plaçons dans ce cadre et, pour définir ce qu'est la "meilleure" réponse, il est nécessaire d'étudier les textes du Web afin de :

- identifier les problèmes auxquels un système question-réponse peut être confronté pour trouver une réponse,
- proposer des méthodes de résolution de ces problèmes.

Nous commençons donc par présenter comment nous avons constitué un corpus d'étude à partir du Web puis nous montrons comment l'analyse de ce corpus nous a permis d'identifier les problèmes, en particulier de quantité et de qualité des données.

1. Constitution du corpus

Un système question-réponse doit proposer à l'utilisateur une réponse correcte à partir d'un ensemble de pages Web traitant du thème de la requête (pages contenant les mots-clés de la requête) obtenu par l'intermédiaire d'un moteur d'extraction. Pour identifier les types de données à manipuler ainsi que comprendre et résoudre les problèmes auxquels un système est confronté, il est nécessaire d'étudier en détail la forme et le contenu des textes du Web. Pour cela, nous avons constitué, via Google ou le système question-réponse QRISTAL[1], un corpus de paires question-réponses (en français) issues du Web.

1.1. Le choix des questions

Dans un premier temps, nous nous sommes appuyées sur une typologie classique des questions [Lehnert, 1978] qui distingue les questions atomiques (qui attendent des réponses de type entité) et les questions narratives (qui attendent des réponses de type texte).
Nous avons adopté une méthode centrée sur les besoins des utilisateurs : le choix des questions à poser a été guidé par les FAQ de sites et les inventaires des questions les plus fréquemment posées sur le Web (données par des générateurs de mots-clés comme ceux de Google[2] et d'Overture[3]). Afin d'avoir un corpus le plus diversifié possible, le corpus est ensuite enrichi

[1] www.qristal.fr, Synapse Développement
[2] https://adwords.google.fr/select/KeywordSandbox
[3] http://inventory.overture.com/d/searchinventory/suggestion/?mkt=fr

manuellement, avec par exemple les questions des campagnes TREC ou des questions imaginées, car certains domaines (domaines spécialisés) sont sous-représentés.

Pour les questions atomiques, nous avons sélectionné des questions qui n'attendent qu'une seule réponse possible (par exemple, *Quand est mort Beethoven ? le 26 mars 1827*) et des questions qui en acceptent plusieurs (par exemple, *Où se trouve le parc Disneyland ? à Paris, Los Angeles, Tokyo, …*). Le tableau 1 donne la distribution des questions atomiques de notre corpus selon le type de réponse attendue.

Type des réponses	**localisation**	**personne**	**numérique**	**temps**	**objet**	**TOTAL**
Nombre de questions	32	34	47	40	27	180

Tableau 1 : Distribution des questions de type atomique

Pour les questions narratives, nous nous sommes intéressées plus particulièrement aux questions procédurales (ces questions induisent une réponse de type procédural, i.e. une suite d'instructions visant à atteindre l'objectif présenté dans la question). En effet, plusieurs études (par exemple, [Yin, 2004]) montrent qu'après les questions atomiques, les questions en *Comment ?* sont les plus nombreuses à être posées sur le Web. Cependant, il existe plusieurs emplois différents du pronom *comment* qui n'introduit pas nécessairement une question procédurale. Grâce au corpus de questions, une typologie des questions procédurales a pu être définie [Aouladomar, 2005]. Le tableau 2 donne la distribution des questions procédurales de notre corpus selon les domaines.

Domaines	**Nb de requêtes procédurales**
Communication - Conseil	48
Informatique - Technique	50
Santé - Médecine	9
Recettes	10
Règlements	7
Total	124

Tableau 2 : Distribution des questions de type procédural

1.2. Le choix des réponses

Les questions sont ensuite soumises à Google sous forme de mots-clés ou à QRISTAL sous forme de question en langue naturelle. Ces deux systèmes fournissent un ensemble de pages comme réponses potentielles. Un premier travail manuel consiste à ne garder parmi cet ensemble que les pages pertinentes qui proposent effectivement une réponse, même fausse, à la question : seules sont conservées les pages qui contiennent le focus de la question et :

- une information correspondant au type sémantique de la réponse attendue pour les questions atomiques,
- des textes dits procéduraux pour les questions procédurales.

Par exemple, la page qui propose comme réponse *Ludwig von **Beethoven** est bien **mort** des suites d'un empoisonnement au plomb* à la question *Quand est mort Beethoven ?* n'est pas conservée car elle ne donne pas d'information temporelle sur la mort de Beethoven. De même, à la question *Comment réparer une fuite d'eau ?*, la page qui propose comme réponse *Vérifiez régulièrement les toilettes, les tuyaux et les robinets pour voir s'il y a des **fuites** et faites immédiatement les **réparations*** n'est pas conservée car elle ne répond pas à la question posée.
Pour la sélection des réponses, nous n'avons considéré que les 20 premiers liens donnés par Google et QRISTAL, les liens suivants n'apportant pas de réponses pertinentes supplémentaires. De plus, cette démarche est cohérente avec les habitudes des internautes qui, pour 80% d'entre eux, ne consultent pas plus d'une dizaine de liens pour une requête[4].

[4] http://www.revue-referencement.com

2. Exploitation du corpus

Une première observation permet de remarquer que la quantité et la qualité des réponses obtenues posent des problèmes pour le choix de la réponse finale. En effet, pour une même question, les réponses potentielles obtenues peuvent être au mieux redondantes, sémantiquement équivalentes mais aussi incohérentes, approximatives, etc. L'analyse du corpus nous permet donc d'identifier, de quantifier et de typer ces problèmes pour pouvoir proposer des solutions au système.

Pour les questions atomiques, l'étude de corpus nous a ainsi permis d'identifier les principales relations pouvant exister entre un ensemble de réponses potentielles. Nous nous sommes inspirées des quatre relations définies dans [Webber et al, 2002], à savoir : l'*équivalence*, l'*inclusion*, l'*agrégation* et l'*alternative* auxquelles nous avons ajouté la relation de *complémentarité* découverte dans le corpus [Moriceau, 2005]. Le tableau 3 montre la distribution des relations entre réponses en fonction du type de réponse attendue.

Type des réponses	Nombre de questions	équivalence	inclusion	agrégation	alternative	complémentarité
localisation	32	11	11	13	3	5
personne	34	13	0	5	11	0
numérique	47	11	0	19	32	16
temps	40	3	0	5	13	5
objet	27	8	3	8	5	3
TOTAL	180	46	14	50	64	29

Tableau 3 : Distribution des relations entre réponses potentielles

De la même manière, les réponses procédurales peuvent être plus ou moins bien construites, destinées à différents publics (grand public / expert), etc. L'avantage de travailler sur un corpus de textes procéduraux issus du Web par rapport aux autres sources textuelles (procédures papiers, livres, etc.), réside dans la possibilité d'obtenir une plus grande diversité des textes tant au niveau de leur source que du type de public visé. Cela nous permet ainsi de récupérer non seulement les textes procéduraux rédigés pour le Web mais aussi ceux initialement diffusés sur support papier et qui sont ensuite mis en ligne. Par ailleurs, travailler uniquement sur un corpus-papier pose le problème de la longueur des réponses : les procédures papiers sont souvent plus longues et trop détaillées avec beaucoup d'informations additionnelles.
Pour notre projet, une analyse fine de la structure des documents procéduraux est nécessaire pour repérer les zones qui serviront pour la formulation d'une réponse appropriée. L'analyse des textes procéduraux permet ainsi de décrire leur structure syntaxique et sémantique et de repérer les différents éléments informationnels présents ainsi que leur organisation et enfin, les différentes marques permettant leur identification. Une première observation du corpus permet de décrire les textes procéduraux comme présentant une structure assez stéréotypée. En effet, on peut observer certaines régularités de contenu : la présence, par exemple, d'un titre présentant l'objectif général à atteindre, d'un groupe d'objets énumérés représentant des pré-requis, et d'une série d'instructions nécessaires pour atteindre ce but. Cette étude nous a permis de définir une grammaire des textes procéduraux [Aouladomar, 2005] qui permet de les annoter.

A partir d'un ensemble de réponses potentielles pour une même question, le système question-réponse doit proposer une réponse correcte. L'étude de corpus a montré que privilégier la 1ère réponse donnée par le moteur de recherche n'est pas une solution acceptable : en effet, Google propose la page qui donne une réponse correcte (identique à celle d'une encyclopédie pour les questions atomiques) en moyenne au 4ème ou 5ème lien (ce chiffre varie bien sûr en fonction du type de question). Il faut donc définir des méthodes qui élaborent une

réponse appropriée en prenant en compte des critères mis en évidence lors de l'étude de corpus tels que les relations que les réponses ont entre elles, leur cohérence, leur structure, leur degré de précision, etc.
Ainsi, pour les réponses atomiques et en particulier les réponses temporelles (dates, intervalles), nous avons défini une méthode d'intégration des réponses potentielles qui tient compte de leur degré de cohérence, de leur niveau de précision, de leur fréquence, de l'itérativité de certains événements, etc. [Moriceau, 2005].
Pour les réponses procédurales, nous avons défini la notion de *questionnabilité* d'un texte procédural : elle exprime la capacité d'un texte à répondre à une question en *comment*. L'analyse des textes de notre corpus nous a ainsi permis de définir une métrique de questionnabilité [Aouladomar, 2005] qui permet dans un premier temps d'identifier les textes procéduraux parmi l'ensemble des pages trouvées par le moteur de recherche puis de comparer ces textes procéduraux afin de retenir le meilleur texte candidat à l'unification avec la question. Le meilleur texte est également sélectionné pour son degré d'informativité, de lisibilité, de précision, de niveau de détail, d'illustrations, etc. Les zones de textes pertinentes sont ensuite extraites et serviront à la formulation finale de la réponse.

3. Validation

Il faut ensuite constituer un autre corpus de questions à soumettre à notre système afin de l'évaluer (pour vérifier qu'il propose bien une réponse correcte). Par exemple, pour l'évaluation du système sur des questions temporelles, nous avons collecté un ensemble de 72 questions issues de la campagne TREC et des inventaires des requêtes temporelles les plus posées sur le Web (celles contenant *quand* ou *date*). Ces questions portent sur différents types d'événements : uniques, itératifs, ponctuels, duratifs (cf. tableau 4). Les résultats de l'évaluation sont présentés dans [Moriceau, 2005].

Réponse attendue	Evénement unique	Evénement itératif	Total
type *point* (ponctuel)	18	18	36
type *intervalle* (duratif)	19	17	36
Total	37	35	72

Tableau 4 : Nombre de questions pour chaque type d'événements évalués

Pour évaluer notre grammaire des textes procéduraux ainsi que les métriques de questionnabilité, nous les avons implémentées en Prolog. Il s'agit d'évaluer si des aménagements ou des affinements de notre grammaire et de nos métriques sont nécessaires ou non. Pour cela, nous avons constitué un corpus plus important de textes procéduraux et non procéduraux (pour tester la questionnabilité) toujours issus du Web. Les résultats sont en cours d'analyse.

Références

[Aouladomar, 2005]. F. Aouladomar, *Towards Answering Procedural Questions*, KRAQ05-IJCAI, Edinburgh, 2005.

[Lehnert, 1978]. W. Lehnert, *The Process of Question Answering: a Computer Simulation of Cognition*, L. Erlbaum, 1978.

[Moriceau, 2005]. V. Moriceau, *Answer Generation with Temporal Data Integration*, ENLG, Aberdeen, 2005.

[Webber et al, 2002]. B. Webber, C. Gardent and J. Bos, *Position statement: Inference in Question Answering*, LREC, 2002.

[Yin, 2004]. L. Yin, *Topic Analysis and Answering Procedural Questions*, Information Technology Research Institute Technical Report Series, ITRI-04-14, University of Brighton, UK, 2004.