Génération de réponses pour un système de questions-réponses

Van-Minh Pho[1]

LIMSI-CNRS, Rue John von Neumann, Université Paris-Sud, 91403 Orsay cedex, France

vanminh.pho@gmail.com

RÉSUMÉ. *Les systèmes de questions-réponses (SQR) actuels répondent à une question posée par l'utilisateur en langue naturelle par une information précise ainsi qu'un passage de document justifiant cette information. Le principal défaut de ces réponses est qu'elles ne créent pas d'interaction avec l'utilisateur, ce qui peut être problématique, en particulier dans le cas où le SQR est intégré dans un système de dialogue oral. Cet article présente un système générant des réponses en langue naturelle et pouvant être intégré à tout SQR.*

ABSTRACT. *Question answering (QA) is the task of automatically answering a question asked in natural language. The answers returned by QA systems contain minimal answers and excerpts of the documents justifying them. The main drawback is that there is no interaction with the user. This paper presents a system for generating natural language answers, which can be embedded in any QA system.*

MOTS-CLÉS : *Interaction avec l'utilisateur, Systèmes de Questions-Réponses, Traitement Automatique de la Langue Naturelle pour la Recherche d'Information, Génération Automatique*

KEYWORDS: *User interface, Question answering systems, Natural Language Processing for Information Retrieval, Natural Language Generation*

1. Ce travail a été encadré par Anne-Laure LIGOZAT et Anne GARCIA-FERNANDEZ.

JJC 2012, pp. 449–454, Bordeaux, 21-23 mars 2012

1. Introduction

Les systèmes de questions-réponses (SQR) ont pour fonction de répondre à des questions formulées en langue naturelle, contrairement aux moteurs de recherche où l'utilisateur saisit un certain nombre de mots-clés correspondant à sa recherche. Une autre différence avec les moteurs de recherche est le résultat de la requête : alors qu'un moteur de recherche retourne un ensemble de pages correspondant aux mots-clés de la requête, les SQR renvoient la réponse précise à la question, ainsi qu'un passage de document la justifiant. L'exemple suivant montre une telle réponse.

> **Question :** *Quelle est la taille de la Tour Eiffel ?*
> **Réponse :** *325 m*
> **Justification :** *D'une hauteur de 313,2 mètres à l'origine, la tour Eiffel est restée le monument le plus élevé du monde pendant 41 ans. Sa hauteur a été augmentée depuis pour culminer à 325 m depuis le 17 janvier 2005.*

Depuis peu, la formulation des réponses en langue naturelle a reçu plus d'attention, permettant ainsi la création de SQR interactifs (Mendes *et al.,* 2004). Dans un dialogue, les interlocuteurs à la recherche d'informations ne souhaitent pas forcément connaître la source, ni les justifications des réponses, mais plutôt obtenir une réponse en langage naturel favorisant ainsi une interaction entre eux. Notre objectif est de développer un système de génération de réponses en langue naturelle, qui prend en entrée une question et une réponse courte fournie par n'importe quel SQR.

2. État de l'art

Ces dernières années, la génération de questions a fait l'objet de nombreux travaux, notamment dans le cadre des campagnes Question Generation. Les méthodes les plus utilisées pour cette tâche sont la génération fondée sur les patrons (Wyse *et al.,* 2009, Curto *et al.,* 2011) et sur la reformulation de la réponse (Kalady *et al.,* 2010).

Un exemple de système de génération fondée sur les patrons est celui de (Wyse *et al.,* 2009), qui prend en entrée des arbres syntaxiques et utilise `Tregex` pour déterminer la compatibilité des phrases avec des règles spécifiques. Ces règles sont associées à des patrons, remplis par les informations extraites de la phrase générant ainsi la question. Par exemple, le patron *«What did X Y ->PAST as ?»* permet de générer des questions telles que *«What did Emmanuel-Joseph Sieyès train as ?»*, où X est un groupe nominal et Y le verbe, à l'infinitif.

Contrairement à la méthode de (Wyse *et al.,* 2009), celle de (Kalady *et al.,* 2010) est fondée sur une reformulation de la phrase complète. Celle-ci subit tout d'abord des opérations de pré-traitement, telles que l'analyse syntaxique, l'extraction des clauses indépendantes et la résolution des anaphores. Pour la génération des questions, plusieurs autres opérations sont effectuées sur la phrase : l'identification de l'interrogatif à utiliser, l'inversion du sujet et du verbe, la suppression des appositions et des groupes prépositionnels.

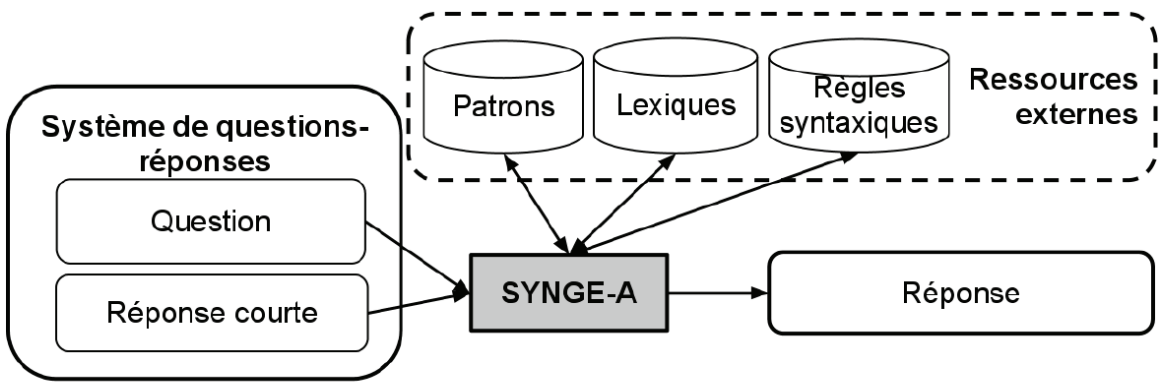

Figure 1. *Génération de réponses par SYNGE-A.*

Il est à noter cependant que la tâche comparable de génération de réponses, à laquelle nous nous intéressons dans cet article, n'a pas été étudiée à notre connaissance.

3. Génération de réponses en langue naturelle

Le système de génération de réponses présenté dans cet article, SYNGE-A (SYstem for Natural GEneration of Answers), prend en entrée une question et une réponse courte retournée par un SQR («*En quelle année a eu lieu l'armistice de la première guerre mondiale ?*», «*1918*»), afin de générer des énoncés réponses.

Il a la particularité d'utiliser les deux méthodes de génération : la génération fondée sur des patrons, et celle fondée sur des règles grammaticales. L'objectif étant d'avoir un système facilement paramétrable et adapté à tous les SQR, SYNGE-A génère un maximum d'énoncés répondant à une question.

3.1. *Architecture*

SYNGE-A prend en entrée un ensemble de couples question-réponse. Si une analyse de la question a été préalablement effectuée, la question a été annotée et la génération des énoncés est fondée sur les patrons (les différentes étapes de cette génération sont détaillées dans la sous-section 3.2). Si ce n'est pas le cas, la question est analysée syntaxiquement afin d'obtenir un arbre syntaxique, permettant ainsi une génération fondée sur les règles grammaticales (la sous-section 3.3 détaille ce processus).

Dans le but de varier les énoncés générés, SYNGE-A utilise un lexique contenant différentes expressions prédéfinies. Ce lexique permet de générer des énoncés génériques, tels que l'énoncé «*Il s'agit d'environ 3,2 kilos*», généré à partir du couple question-réponse («*Combien pèse un bébé à la naissance ?*», «*environ 3,2 kilos*»).

Notre objectif étant d'avoir un système générique, toutes ses ressources (patrons, règles grammaticales et lexique) sont externalisées. Cela permet de ne pas limiter

l'utilisation de SYNGE-A à la langue française et d'utiliser d'autres ressources dans le cas où il serait utilisé pour une autre langue.

3.2. *Générateur fondé sur les patrons*

L'utilisation de patrons est la méthode la plus simple pour générer des énoncés. Les patrons sont composés de trois types d'éléments : des éléments de la question, qui vont être instanciés en fonction de l'analyse de la question, des expressions prédéfinies, et la réponse. Ainsi, pour la question «*Dans quelle ville se trouve Le Louvre ?*», la question est analysée de la façon suivante : «*Dans quelle <type attendu>ville</type attendu> <verbe principal>se trouve</verbe principal> <objet>Le Louvre</objet> ?*». Une réponse possible est «*Paris*», et un patron générant un énoncé réponse est le suivant : «*<type attendu> où <verbe principal> <objet> est <information-réponse>*», ce qui génère l'énoncé «*La ville où se trouve le Louvre est Paris.*»

3.3. *Générateur fondé sur les règles grammaticales*

Le générateur de SYNGE-A fondé sur les règles grammaticales prend en entrée une analyse syntaxique du couple question-réponse, effectuée par le `Bonsai Berkeley Parser` (Candito *et al.,* 2010). Afin de générer les énoncés répondant à la question, l'arbre syntaxique de la question est modifié à l'aide des outils `Tregex` et `Tsurgeon` (Levy *et al.,* 2006). `Tregex` permet d'effectuer des recherches sur les arbres et `Tsurgeon` effectue la transformation des arbres sélectionnés par `Tregex`. Les règles de génération des énoncés sont un ensemble d'opérations de manipulation d'arbres (déplacement, insertion et suppression de nœuds).

Prenons comme exemple le couple question-réponse («*Combien pèse une bouteille d'eau ?*», «*2 kilos*»). L'analyse syntaxique de la question indique que «*combien*» est un adverbe interrogatif, «*pèse*» est un groupe verbal et «*une bouteille d'eau*» est un groupe nominal. Une règle applicable est la suivante, qui contient trois opérations : la première est la suppression de l'adverbe interrogatif, la seconde est le déplacement du groupe nominal à gauche du groupe verbal et la dernière est l'insertion de l'information-réponse à droite du groupe verbal. Cette règle permet la génération de l'énoncé «*Une bouteille d'eau pèse 2 kilos.*».

3.4. *Ajustements linguistiques de surface*

La génération seule ne suffit pas à produire des énoncés linguistiquement corrects. Par exemple, l'énoncé «*La Joconde se trouve à le Louvre*» est grammaticalement correct, mais ne l'est pas morphologiquement : la préposition «*à*» suivie par le déterminant «*le*» est incorrecte en français, ces mots devant être contractés afin d'obtenir le mot «*au*». C'est pour cette raison qu'une étape d'ajustement linguistique est obligatoire, dans le but d'obtenir des énoncés en langue naturelle. De plus, si SYNGE-A est

intégré à un système de dialogue écrit, des ajustements typographiques doivent être effectués : par exemple, en français, une phrase commence par une majuscule et finit par un point.

4. Réponses générées

SYNGE-A a été testé à partir du corpus MACAQ, corpus multi-annoté de réponses à des questions, en langue française (Garcia-Fernandez *et al.,* 2010). 63 patrons et 20 règles grammaticales du français ont été écrites. La raison pour laquelle moins de règles grammaticales sont écrites est la difficulté de varier les énoncés : en effet, les règles dépendent uniquement de la structure syntaxique des questions.

Énoncés générés par le générateur fondé sur les patrons
Il s'agit d'environ 3,2 kilos.
Environ 3,2 kilos.
Un bébé pèse environ 3,2 kilos à la naissance.
Énoncés générés par le générateur fondé sur les règles grammaticales
Un bébé pèse environ 3,2 kilos à la naissance.
Un bébé pèse environ 3,2 kilos.

Tableau 1. *Exemple d'énoncés générés à partir du couple question-réponse («Un bébé pèse combien à la naissance ?», «environ 3,2 kilos»).*

Le tableau 1 montre différents énoncés produits par les générateurs fondés sur les patrons et les règles grammaticales. 13 énoncés ont été générés à partir du générateur fondé sur les patrons, tandis que 5 l'ont été à partir du générateur fondé sur les règles grammaticales.

Questions booléennes	30.5
Questions de *quantité*	11.5
Questions de *lieu*	10.5
Questions de *temps*	9

Tableau 2. *Nombre moyen d'énoncés générés en fonction de la catégorie de la question.*

En utilisant l'ensemble du corpus MACAQ (707 questions), le tableau 2 indique le nombre moyen d'énoncés générés par question en fonction de chaque catégorie. À partir de ce corpus, 2681 énoncés non triviaux (contenant d'autres éléments que l'information-réponse) ont été générés avec le générateur fondé sur les patrons. Nous avons évalué manuellement ces énoncés : 2365 (88 %) sont syntaxiquement et sémantiquement corrects.

5. Conclusion

SYNGE-A est un système générant des énoncés en utilisant deux méthodes différentes : la génération fondée sur les patrons, et celle fondée sur les règles grammaticales. Conformément aux contraintes définies dans cet article, toutes les ressources de SYNGE-A ont été externalisées, ce qui lui assure une grande généricité. Ce système doit encore être testé avec d'autres corpus de questions et pour d'autres langues que le français.

D'autre part, les générateurs de SYNGE-A peuvent encore être améliorés. En particulier, pour le générateur fondé sur les règles grammaticales, il serait intéressant de procéder à une simplification de la question en éliminant certains compléments.

6. Bibliographie

Candito M., Nivre J., Denis P., Henestroza Anguiano E., « Benchmarking of statistical dependency parsers for French », *COLING*, ACL, p. 108-116, 2010.

Curto S., Mendes A. C., Coheur L., « Exploring linguistically-rich patterns for question generation », *Proceedings of the UCNLG+Eval : Language Generation and Evaluation Workshop*, ACL, p. 33-38, 2011.

Garcia-Fernandez A., Rosset S., Vilnat A., « MACAQ : A Multi Annotated Corpus to study how we adapt Answers to various Questions », *LREC*, 2010.

Kalady S., Elikkottil A., Das R., « Natural language question generation using syntax and keywords », *Proceedings of QG2010 : The Third Workshop on Question Generation*, p. 1-10, 2010.

Levy R., Andrew G., « Tregex and Tsurgeon : tools for querying and manipulating tree data structures », *LREC*, p. 2231-2234, 2006.

Mendes S., Moriceau V., « L´analyse des questions : intérêts pour la génération des réponses », *Workshop Question-Réponse*, 2004.

Wyse B., Piwek P., « Generating Questions from OpenLearn study units », *Proceedings of the International Conference on Artificial Intelligence in Education Workshops (AIED)*, 2009.