De l'importance des synonymes pour la sélection de passages en question-réponse

**Brigitte Grau[*] — Anne-Laure Ligozat[*] — Isabelle Robba[*]
Anne Vilnat[*] — Faïza El Kateb[*] — Gabriel Illouz[*] — Laura Monceaux[**]
Patrick Paroubek[*] — Olivier Pons[***]**

[*] *LIMSI-CNRS, BP 133 91403 Orsay Cedex*

prénom.nom@limsi.fr

[**] *LINA, 2 rue de la Houssinière,BP 92208, 44322 Nantes Cedex 03*

Laura.Monceaux@lina.univ-nantes.fr

[***] *CEDRIC-IIE, 18 allee Jean Rostand, 91025 Evry Cedex*

pons@cnam.fr

RÉSUMÉ. *Les systèmes de question-réponse développés actuellement adoptent pour la plupart et à peu de chose près le même type d'architecture que l'on peut schématiser en trois modules : l'analyse de la question, la sélection des documents, l'extraction de la réponse. Mais ce en quoi ils diffèrent, ce sont les outils (moteur d'indexation, analyseurs...) et les bases de connaissances qu'ils utilisent. Pour chacun de ces systèmes, il est donc important d'évaluer l'apport de ces outils ou bases de connaissances. Dans le cadre de la campagne Equer (campagne d'évaluation des systèmes de question-réponse pour le français), notre système FRASQUES a produit deux jeux de résultats : l'un utilise des synonymes dans les bi-termes, l'autre pour les mono-termes aussi. La comparaison de ces deux tests et l'étude d'un corpus plus large, en français et en anglais, permet de mesurer l'apport de ces connaissances sémantiques.*

ABSTRACT. *Most of the question answering systems currently developped adopt a fairly similar architecture, which can be divided into three modules : question analysis, document retrieval and answer extraction. However they differ in their tools (indexing engine, parsers...) and the knowledge bases they use. Thus, for each of these systems, it is important to estimate the contribution of these tools or knowledge bases. In the context of the Equer campaign (evalution campaign for French question answering systems), our system FRASQUES produced two runs : one used synonyms for bi-terms only, the other for mono-terms too. The comparison of these two tests and the study of a broader corpus, in French and in English, allow us to measure the contribution of this kind of semantic knowledge.*

MOTS-CLÉS : *Système de question-réponse, synonymie, évaluation*

KEYWORDS: *Question-answering system, synonymy, evaluation*

1. Introduction

Explorer une liste de documents pour trouver une information précise n'est pas toujours chose facile. Le domaine de la recherche d'information a en effet pour objectif principal d'informer l'utilisateur sur un sujet particulier, mais ne permet pas nécessairement de trouver une réponse précise à une question factuelle. C'est à ce besoin que se proposent de répondre les systèmes de question-réponse (QR). À une question comme "Qui a fondé le Festival international de la bande dessinée d'Angoulême ?", "Combien y a-t-il d'habitants à Saint-Chéron ?" ou "À quelle société appartient le satellite Asisat-2 ?", un tel système devra répondre par une réponse courte, au lieu d'une liste de documents à parcourir.

Les systèmes de QR se situent donc à la frontière de plusieurs domaines de recherche, parmi lesquels la recherche d'information et le traitement automatique des langues jouent des rôles majeurs. Avec l'essor d'Internet, les interrogations à l'aide d'un moteur de recherche sont de plus en plus nombreuses. Mais, si les systèmes de QR étaient complètement opérationnels, ils remplaceraient en partie ces moteurs et ce parce qu'ils fournissent une réponse précise et exacte et, dans une moindre mesure, parce qu'ils permettent d'exprimer les requêtes en langage naturel.

Pour cette raison, les systèmes de QR sont aujourd'hui un enjeu important, comme l'ont compris les organisateurs des campagnes d'évaluation américaines TREC[1] qui ont inclus depuis 1999 la tâche de question-réponse à leurs évaluations, ou ceux des campagnes européennes CLEF[2], où la tâche de question-réponse multilingue est présente depuis 2003. Ces évaluations suivent le même principe. Les systèmes doivent extraire d'une grande collection de documents la réponse à un ensemble de questions portant sur tout domaine. Il s'agit de questions dont la réponse tient en peu de termes comme le nom d'une personne, d'un événement, d'une entité ou d'une caractéristique. Cela peut aussi être une manière, une courte explication. Le but est donc de fournir seulement la réponse, et non un extrait contenant la réponse. Nous avons retenu pour notre étude le corpus de l'ensemble des passages réponses, bons et mauvais, fournis par les participants des campagnes TREC8 et TREC9.

En été 2004, la campagne Equer[3] a permis à sept systèmes français de comparer leurs performances sur la tâche générale : 3 laboratoires de recherche, 2 entreprises, 2 organismes de recherche et développement. De cette évaluation spécifique au français, nous avons retenu deux corpus, le corpus de l'ensemble des bonnes réponses de tous les participants, et notre corpus de réponses (bonnes et mauvaises).

L'une des difficultés attendues dans une tâche de question-réponse est la distance sémantique entre les mots de la question et des passages réponses présents dans les documents. En effet, les formulations des questions dans les évaluations sont *a priori*

1. Organisées par le NIST : http ://trec.nist.gov
2. http ://clef.isti.cnr.it/
3. La campagne Equer fait partie du projet Technolangue(http ://www.technolangue.net) et a été organisée par Elda (http ://www.elda.org/article118.html)

indépendantes des phrases du corpus, et peuvent donc différer de manière significative. La prise en compte de ces variations peut se faire soit au moment de la requête et dans ce cas cela demande de lever l'ambiguïté des mots, afin de ne pas engendrer une expansion trop large, soit au moment de la sélection des passages, en ayant veillé à ce que la requête ne soit pas trop restrictive de manière à ce que les documents pertinents aient été retrouvés. Notre système de QR, lui, tient compte des reformulations sémantiques des termes, simples ou composés, de la question afin de restreindre l'ensemble des documents résultats du moteur de recherche ; mais la pertinence et les performances de cette stratégie n'avaient pas encore été complètement évaluées. Il nous a donc paru intéressant de mesurer sur les corpus dont nous disposons désormais sur le français, l'importance de ce problème pour notre système de QR et plus généralement pour tout système de QR. Le but de cet article est aussi de montrer dans quelle mesure un système de question-réponse peut trouver des réponses sans presque utiliser de ressources sémantiques et quelle est la part des performances qui pourraient être améliorée par de telles ressources. Afin de valider nos résultats, nous les comparerons à ceux obtenus sur de l'anglais.

Nous présenterons en premier lieu, l'architecture de FRASQUES[4], notre système de QR, puis nous montrerons à travers quelques systèmes comment les connaissances sémantiques sont utilisées en QR. Nous décrirons ensuite comment les synonymes sont utilisés dans FRASQUES. Enfin, nous donnerons un certain nombre d'évaluations sur les corpus étudiés.

2. Architecture du système FRASQUES

Comme la plupart des systèmes de QR, FRASQUES s'organise en trois principaux modules présentés dans la figure 1 : l'analyse des questions, la sélection des documents par un moteur de recherche, et le traitement des documents pour en extraire les passages et les réponses finales. FRASQUES a été réalisé en suivant les mêmes grands principes que ceux de QALC ([FER 02]), notre système sur l'anglais, même si dans la réalisation les deux systèmes diffèrent.

L'analyse des questions est réalisée en 2 étapes. L'analyseur XIP de Xerox ([AïT 02]) construit les segments syntaxiques et établit les relations entre eux. À partir de ces données sont calculées des informations telles que le focus, i.e. l'objet à propos duquel une information est demandée, ou le type attendu de la réponse, quand il s'agit d'un type appartenant à notre liste d'entités nommées. C'est également au cours de l'analyse de la question que sont construites les listes de synonymes des mots non vides de la question (nous y reviendrons au paragraphe 4). Le moteur de recherche utilisé, Lucene[5], est un moteur booléen ; il nous a permis d'indexer le corpus (qui avait été au préalable lemmatisé à l'aide du Tree Tagger[6] et de l'analyseur morphologique de XIP). Lors de l'interrogation, Lucene reçoit un ensemble de requêtes constituées des mots non vides

4. FRench Answering System to QUEStions
5. http ://jakarta.apache.org/lucene/docs/index.html
6. http ://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html

de la question. Si le moteur ne retourne aucun document (ou un nombre inférieur à un seuil), nous réinterrogeons la base avec moins de termes à chaque fois (voir le détail au paragraphe 5.3). Les documents trouvés par Lucene sont re-indexés par Fastr ([JAC 96]) qui permet de reconnaître des variantes morphologiques, syntaxiques ou sémantiques des termes simples et composés de la question. Ceux-ci sont pondérés et permettent de donner un poids aux documents. Les documents sont ainsi réordonnés et un sous-ensemble est extrait sur lequel on applique alors le module étiquetant les entités nommées. Enfin le module d'extraction de la réponse est appppliqué. Il procède différemment si la question attend ou non pour réponse une entité nommée. Dans FRASQUES, comme dans la plupart des systèmes de QR, les questions attendant en réponse une entité nommée obtiennent de meilleurs résultats que les autres, car, de par leur nature, les entités nommées sont plus facilement repérées dans les documents.

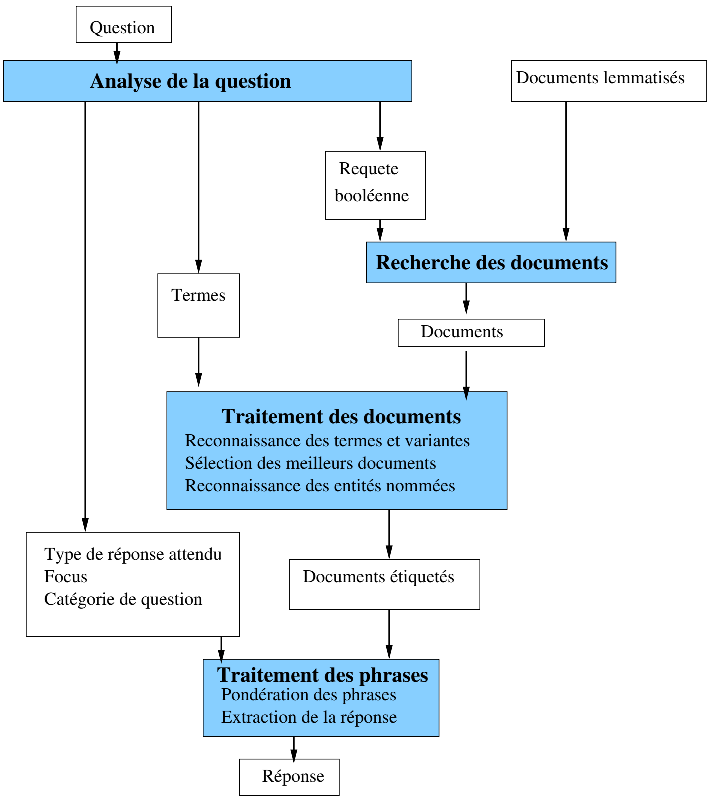

Figure 1. *Architecture du système FRASQUES*

3. L'utilisation de la sémantique pour la sélection des documents

Pour améliorer la sélection des documents ou des passages dans un système de QR, une stratégie possible consiste à prendre en compte un certain nombre de connaissances sémantiques, extraites le plus souvent de thesauri ou de lexiques. Ces connaissances peuvent être utilisées à plusieurs niveaux de la chaîne de traitement : soit directement au niveau de la requête fournie au moteur de recherche, soit ultérieurement au niveau de la sélection des documents les plus pertinents. On retrouve là les deux approches citées par [MOR 05]. On peut toutefois noter que le problème se pose de manière différente dans un système de question-réponse et dans un système de recherche de documents. Ces derniers cherchent à répondre à des requêtes thématiques, et il est important de favoriser le rappel et la précision des résultats, soit de retrouver tous les documents relatifs à un sujet, et au mieux seulement ceux-là. Dans un système de question-réponse, le but est de trouver la réponse, et non tous les documents contenant la réponse. C'est pourquoi les système de QR favorisent en général la précision.

La différence dans les buts de la recherche se retrouve aussi dans les techniques de gestion de la variation sémantique. Les système de QR utilisent principalement des thésauri (WordNet en général), permettant de sélectionner des extensions en fonction des relations entre mots ou concepts, par rapport à l'utilisation de proximités sémantiques qui conduisent à des extensions thématiques qui sont en général moins pertinentes dans le cas de la recherche de réponses précises.

Au niveau de la construction de la requête, il est tout d'abord nécessaire de bien choisir les mots-clefs à utiliser, avant d'étendre éventuellement la requête par utilisation de connaissances lexicales. La sélection des mots-clefs se fait généralement en fonction de la catégorie morpho-syntaxique des mots de la question. Ces mots-clefs peuvent ensuite être répartis en différents groupes auxquels on attribuera des poids ou des rôles différents dans la requête, soit en fonction de règles préalablement établies, soit en fonction de poids calculés à partir d'un corpus.

[MOL 03] proposent trois facteurs de sélection des mots-clefs, qu'il faudrait idéalement utiliser dans une heuristique de sélection des mots-clefs : la saillance sémantique, la redondance et le degré de variation. Ces facteurs étant difficiles à déterminer sur des questions en domaine ouvert, 3 catégories de mots ont été définies, qui tentent de tenir compte de ces facteurs :

– les termes de haute pertinence : citations, comparatifs et superlatifs et noms propres ;

– les termes de pertinence moyenne : noms communs à l'exception des termes permettant de déterminer le type de la réponse ;

– les termes de faible pertinence : verbes et noms communs désignant le type de la réponse.

Ces termes sont ensuite utilisés pour construire une requête booléenne de taille variable.

[LAV 04] modifient quant à eux le modèle vectoriel classique de recherche d'information pour l'adapter à leur système de QR. Après avoir étudié sur un corpus de TREC-9 la répartition des mots-clefs dans les phrases réponses en fonction de leur catégorie morpho-syntaxique, ils proposent d'attribuer des poids plus élevés aux catégories les mieux représentées dans les phrases réponses.

En plus des mots-clefs, il peut être intéressant d'utiliser un certain nombre de leurs reformulations, afin de tenir compte de l'éventuelle distance lexicale entre les questions et les phrases réponses.

[MOL 03] tirent ainsi parti de WordNet pour générer des variations morphologiques, lexicales et sémantiques des mots-clefs, et utilisent ces termes pour étendre progressivement les requêtes si une partie de leur système ne retourne pas assez de réponses.

Le système de [YAN 02] intègre les deux ressources externes que sont le Web et WordNet pour étendre leur requête : après avoir lancé la question sur un moteur de recherche sur le Web, ils retiennent les termes qui sont fortement corrélés aux mots-clefs de la question, et les considèrent comme des termes possibles de la requête, puisque leur association aux mots-clefs de la question semble pertinente dans le contexte de la question ; d'autre part, certains termes reliés dans WordNet aux mots de la question (par des relations de définition, de synonymie ou d'hyperonymie) sont ajoutés.

Ittycheriah et al. ([ITT 01]) ont testé différentes interrogations, avec et sans expansion de requête, afin de mesurer si il était préférable d'étendre les requêtes ou de filtrer les documents après une première sélection. Ils ont trouvé des résultats meilleurs quand ils appliquent les mécanismes d'expansion comme critères de sélection de la réponse dans les documents retrouvés par le moteur plutôt que donner au moteur des requêtes étendues. C'est aussi le choix fait dans FRASQUES.

Aucun système, à notre connaissance, n'a réellement mesuré l'apport des connaissances sémantiques qu'il utilise, ni l'apport que l'on est en mesure d'attendre de telles connaissances pour répondre à des questions. C'est pourquoi cet article tente d'y apporter une première réponse.

4. Les synonymes dans le système FRASQUES

Pour sélectionner les documents et passages pertinents, le moteur de recherche s'appuie sur les mots-clefs de la question, qui sont dans le cas du système FRASQUES, les mots considérés comme non vides, déterminés par leur étiquette morpho-syntaxique (noms, verbes, adjectifs). Mais ces mots ne suffisent pas nécessairement à récupérer le bon passage. Ainsi, à la question "Citez le nom d'une province sénégalaise", une réponse possible est " Celles-ci doivent mettre un terme au conflit armé qui oppose depuis dix-sept ans le MFDC à l'armée sénégalaise en Casamance, la région sud du Sénégal.", dans laquelle le mot plein de la question "province" n'est pas présent. Il est donc crucial de considérer, en plus des mots-clefs de la questions, un certain nombre de leurs reformulations ou synonymes.

Nous disposons de synonymes provenant de deux sources distinctes. La première source est celle utilisée par le système Fastr qui se fonde sur deux bases de connaissances, l'une constituée des familles morphologiques et l'autre de synonymes, pour reconnaître dans les documents les mono- et multi-termes et leurs variantes aussi bien morphologiques, syntaxiques que sémantiques. Par abus de langage, nous les nommerons synonymes Fastr par la suite. L'autre source est la base de connaissances EuroWordNet[7]. Le nombre de synonymes correspondant aux mots de la question est parfois élevé (des chiffres plus précis sont donnés au paragraphe 5.1), et malheureusement nous n'avons disposé pour Equer d'aucun outil nous permettant d'écarter les synonymes inutiles dans le contexte de la question. Ainsi pour la question "Quel est le nom du parti politique de M. Vajpayee ?" les synonymes proposés par Fastr pour le mot "parti" sont : *groupe*, *avantage*, *décision*, *fiancé*, *éméché*. EuroWordNet, plus laconique, ne propose que *parti politique*.

Cet exemple montre qu'il faut utiliser des moyens de désambiguisation. Cependant dans le cas des questions, des outils reposant sur les autres mots du contexte ne seraient bien souvent pas suffisants. Ainsi dans le cas d'une question telle que "Où est situé Belfast ?", les synonymes proposés par Fastr pour le verbe "situer" sont : *placer*, *déceler*, *siéger* ; et ceux proposés par EuroWordNet : *localiser*, *placer*, *identifier*, *reconnaître*. Pour effectuer la désambiguisation qui nous permettrait de ne garder que *localiser* et *placer*, il nous faudrait faire une analyse de la question afin de déterminer que l'objet du verbe est un nom de lieu, que la question porte sur une localisation et donc que dans ce contexte seuls *localiser* et *placer* doivent être conservés. En regard de ces difficultés, nous avons choisi de sélectionner très largement des documents, et de gérer l'ambiguïté à leur niveau plutôt qu'au niveau de la seule question, comme nous le verrons par la suite.

5. Analyse du système FRASQUES

5.1. *Les questions*

Le module d'analyse des questions de FRASQUES détermine pour chaque question, un ensemble d'informations qui sont utilisées par le module d'extraction de la réponse. Parmi ces informations, trois ensembles nous intéressent plus particulièrement dans cet article :

– l'ensemble des mots non vides de la question : les noms, les verbes, les adjectifs et les adverbes sont retenus ;

– l'ensemble de leurs synonymes issus de Fastr ;

7. Nous tenons à remercier Christine Jacquin (du LINA, Nantes) dont les travaux ont permis une interrogation de la base EuroWordNet, et qui nous a transmis les outils nécéssaires à l'extraction de ces synonymes depuis EuroWordNet

– l'ensemble de leurs synonymes issus d'EuroWordNet.

Dans le cadre de la campagne Equer, la tâche générale[8] comportait 500 questions. Parmi ces 500 questions, 33 ne possèdent pas de synonyme Fastr et 73 ne présentent aucun synonyme EuroWordNet. La table 1 nous donne la somme des cardinalités des 3 ensembles pour les 500 questions, et les moyennes par question.

	Nombre total	Moyenne par question
Nombre de mots dans les questions	2815	5.6
Nombre de synonymes Fastr	6422	12.8
Nombre de synonymes EuroWordNet	3560	7.1

Tableau 1. *Quelques chiffres issus de l'analyse des 500 questions*

On constate qu'il y a en moyenne plus de 2 synonymes Fastr par mot de la question, ce qui est relativement élevé, surtout si l'on tient compte du fait que dans les 500 questions d'Equer, on dénombre 592 noms propres et que ceux-ci n'ont pas ou fort rarement de synonyme. Si l'on soustrait ce nombre de noms propres du nombre des mots de la question, cela porte alors à presque 3 le nombre de synonymes Fastr par mot des questions.

On observe aussi qu'il y a presque deux fois plus de synonymes Fastr que de synonymes EuroWordNet, ce qui peut être expliqué par le fait que la base EuroWordNet a réellement une faible couverture dans certains domaines.

5.2. *Etude du corpus des passages corrects des participants*

L'ensemble des passages fournis par les participants à Equer et jugés corrects par Elda, lors de l'évaluation, constitue un corpus qu'il était intéressant d'étudier. Pour mesurer l'apport de connaissances telles que les synonymes, nous avons cherché à savoir si la présence des synonymes dans les passages corrects était déterminante (les passages font au maximum 250 caractères).

Le corpus est constitué de 3010 passages non vides, soit environ 6 passages en moyenne par question. Dans ce corpus seules 5 questions n'obtiennent aucune réponse.

Parmi ces 3010 passages, il est surprenant de voir que 2204 (73%) ne contiennent aucun synonyme Fastr, et 2160 (72%) ne contiennent aucun synonyme EuroWordNet.

Dans ce corpus, seuls les mots de la question ont un taux de présence important comme le montre la table 2 qui contient les taux de présence des mots non vides de la question et des synonymes Fastr et EuroWordNet. De notre point de vue, diverses raisons peuvent expliquer de si faibles taux de présence des synonymes. Tout d'abord

8. La campagne Equer proposait aussi une tâche spécialisée portant sur le domaine médical, mais nous n'y avons pas participé

Pourcentage de synonymes Fastr présents	1.7
Pourcentage de synonymes EuroWordNet présents	2.0
Pourcentage de mots de la question	59.1

Tableau 2. *Mots de la question et synonymes présents dans les passages corrects*

ces bases de synonymes ne sont pas celles utilisées par les autres participants qui sont de plus peu nombreux à utiliser de telles connaissances.

D'autre part, on observe que dans la campagne Equer beaucoup de bonnes réponses pouvaient être obtenues grâce à la présence des mots de la question. On a en fait le sentiment (et cela est vrai aussi pour les campagnes TREC) qu'il existe souvent au moins une formulation proche de la question, ce qui peut être expliqué par la taille importante de la base des documents (1,5 gigaoctets).

Une étude intéressante serait de rechercher sur une collection encore beaucoup plus large (sur Internet par exemple) les réponses à ces questions afin de voir dans quelle mesure les synonymes sont indispensables, et ce sur un corpus qui est complètement dissocié des questions.

Après avoir étudié les distributions des mots-clefs et des synonymes dans le corpus général, nous nous sommes intéressés au corpus des phrases retournées par FRASQUES, pour lequel nous disposons de deux sous-corpus : l'ensemble des phrases jugées correctes, et l'ensemble des phrases jugées incorrectes. L'étude de ces phrases nous a permis d'évaluer la pertinence de notre approche, fondée sur la recherche des mots-clefs et de leurs synonymes.

5.3. *Etude des passages retournés par FRASQUES*

En ce qui concerne le système FRASQUES, nous disposons du corpus des passages retournés, qui comprend notamment l'ensemble des documents renvoyés par le moteur de recherche Lucene, et l'ensemble des documents sélectionnés et ordonnés après leur ré-indexation par Fastr. Partant de ces corpus, nous avons cherché à étudier l'influence de nos connaissances sémantiques sur la sélection des passages, et ceci aux différentes étapes de la chaîne de traitement.

L'interrogation par le moteur a été effectuée de la même manière dans les deux jeux de test. Le but est de favoriser les documents qui contiennent tous les mots de la question, à l'identique. Si de tels documents n'existent pas, ou en trop petit nombre, nous relâchons des contraintes et c'est de cette manière que les mots non présents dans la requête peuvent éventuellement être retrouvés sous une autre forme dans les documents renvoyés.

Nous commençons par constituer une requête composée de tous les mots pleins de la question (noms, adjectifs, verbes, noms propres). Si le nombre limite de documents,

	test1	test2
Après le moteur de recherche	342/450 (76%)	330/450 (73%)
Après Fastr	idem	idem

Tableau 3. *Pourcentage de questions pour lesquelles la réponse est présente dans les documents retournés par FRASQUES*

fixé à 200, n'est pas atteint, nous envoyons une requête constituée du nom focus, du verbe principal et des noms propres. Ensuite, la relaxation consiste à enlever le verbe, puis à constituer des requêtes différentes pour chaque nom propre. Pour test1, tous les noms propres étaient essayés sans tester le dépassement du seuil limite, pour test2 nous avons imposé ce seuil après chaque requête, et ce uniquement pour des questions de temps de calcul.

Par ailleurs, nous avons recensé pour chaque question les différentes réponses données par les participants et l'organisateur, et nous avons testé la présence de ces réponses dans les documents ou passages retournés par notre système. Les résultats pour les premières étapes de notre système, que sont les sélections de documents par le moteur de recherche puis après la sélection à l'issue de Fastr, sont présentés dans la table 3[9].

Le moteur de recherche ne retourne des documents contenant la réponse que pour 73 à 76% des questions. Cela peut s'expliquer par plusieurs facteurs : imprécision de la sélection des mots-clefs des questions, qui sont sélectionnés uniquement en fonction de leur étiquette morpho-syntaxique ; erreurs de lemmatisation (ainsi, dans la question "Comment Dominique Voynet est-elle allée à l'Elysée le 22 septembre 1999 ?", "allée" est étiqueté comme un nom), problèmes de référence... Ces difficultés étaient déjà présentes dans le système de QR pour l'anglais du LIMSI [FER 02].

La sélection de 50 documents opérée après l'indexation par Fastr et fondée sur les reformulations des multi-termes trouvés (et aussi des synonymes mono-termes pour test2) n'entraîne pas de perte de bons documents. Le second test ayant utilisé des synonymes mono-termes, on pourrait s'attendre à ce que son rappel soit meilleur, mais il n'en est rien. Ce phénomène peut s'expliquer d'une part par le fort degré de ressemblance des questions avec les phrases réponses dans l'évaluation Equer, et d'autre part par le bruit introduit par la recherche de "mauvais" synonymes.

Des variantes sémantiques de multi-termes ont été trouvées par Fastr pour 40 questions (9% des questions). Ces reformulations ont permis par exemple d'effectuer des rapprochements entre les expressions "transfert d'animal" et "transport des animaux", ou "avocat de M." et "défenseurs de M.". Les synonymes utilisés ici semblent plus pertinents que ceux utilisés pour les mono-termes, pour lesquels le manque de contexte

9. Le nombre des questions est ici ramené à 450 : les questions booléennes et les questions dont la réponse est une liste n'ayant pas été prises en compte

	Passages jugés corrects	Tous les passages
Mots de la question	69%	56%
Synonymes Fastr	2.2%	2%
Synonymes EuroWordNet	3.1%	2.3%

Tableau 4. *Taux de mots de la question et de synonymes présents dans les phrases retournées par FRASQUES*

ne permet pas de discriminer entre tous les synonymes possibles. Les reformulations multi-termes révèlent donc bien ici leur intérêt, et il pourrait être utile de les favoriser à d'autres niveaux de la chaîne.

Nous avons également mené quelques évaluations sur la pertinence des mots de la question pour la recherche de la réponse, qui sont résumées dans la table 4. Pour chaque phrase réponse retournée par notre système, nous avons calculé le nombre de mots de la question présents, ainsi que le nombre de synonymes Fastr et EuroWordNet, puis nous avons effectué ces mêmes calculs uniquement sur les passages jugés corrects. Il est intéressant de remarquer que les phrases jugées correctes présentent des taux plus élevés de mots de la question et de synonymes, car cela justifie bien notre approche actuelle de sélection de passages.

Les taux de synonymes dans nos phrases réponses sont comparables à ceux trouvés sur le corpus de réponses des participants. On peut également remarquer que, comme le corpus des participants, notre corpus contient un nombre important de phrases ne contenant pas de synonyme : entre 71 et 76 %, en fonction de l'origine de synonymes et des phrases considérées. Cette très faible présence des synonymes est probablement due, comme évoqué précédemment, à l'imprécision du choix des synonymes, et à la proximité lexicale des questions et des phrases réponses.

Afin de vérifier si ces chiffres étaient spécifiques à l'évaluation EQUER ou s'ils avaient quelque généralité, nous avons effectué quelques mesures similaires sur les corpus provenant des campagnes TREC8 et TREC9.

6. Analyse quantitative des corpus de Trec8 et Trec9

Après toute campagne d'évaluation les ressources annotées ou tout au moins évaluées sont très recherchées par les participants qui souhaitent améliorer leur système. Abe Ittycheriah (du centre de recherche d'IBM Watson), qui a participé à Trec8 et Trec9 a mis à disposition de la communauté deux fichiers dans lesquels on trouve les phrases ou les paragraphes contenant les réponses données par tous les participants et issus de la collection[10].

10. Ces fichiers sont disponibles à l'adresse http ://trec.nist.gov/data/qa/add_qaresources.html

Certes cette ressource n'est pas aussi pertinente qu'un corpus constitué uniquement de bonnes réponses. Cependant c'est une ressource de très grande taille : 893 questions et 104 077 passages[11] (après suppression des doublons) et qui plus est en anglais. Soulignons que Trec8 était la première édition de Trec incluant la tâche de question-réponse, comme pour Equer, c'était donc un premier essai pour les participants.

Dans un premier temps près d'un tiers de chaque fichier a été supprimé car il s'agissait de passages en plusieurs exemplaires. Les passages et les questions ont été lemmatisés à l'aide du Tree Tagger, puis pour chaque question, seuls les mots pleins ont été conservés, à l'exception des auxiliaires do, have et be (qui présente un grand nombre de synonymes non pertinents et sont présents dans énormément de questions). L'ensemble des synonymes que nous avons utilisé pour ce corpus en anglais est issu de WordNet.

Les évaluations que nous avons effectuées sur ces deux nouveaux corpus donnent des résultats sensiblement proches de ceux obtenus sur le corpus d'Equer, sauf en ce qui concerne le pourcentage de mots pleins des questions présents dans les passages. Ce qui est bien naturel puisque là nous cummulons bons et mauvais passages. Sur Equer il était de 59.1 %, sur Trec8 et Trec9 il est de l'ordre de de 36 %. Les tables 5 et 6 concernent respectivement les 200 questions de Trec8 et les 693 de Trec9. Le nombre moyen de mots par question est plus faible pour Trec9 (3.48) ce qui est dû au fait que Trec9 contenait beaucoup de questions de définition qui ne comportent qu'un seul mot plein : "Who is Picasso ?" ou "What is platinum ?" en sont deux exemples.

	Nombre total	Moyenne par question
Nombre de mots dans les questions	1017	5.08
Nombre de synonymes WordNet	3261	16.3

Tableau 5. *Quelques chiffres issus de l'analyse des 200 questions de trec8*

	Nombre total	Moyenne par question
Nombre de mots dans les questions	2411	3.48
Nombre de synonymes WordNet	8394	12.11

Tableau 6. *Quelques chiffres issus de l'analyse des 693 questions de trec9*

En ce qui concerne les passages (tables 7 et 8), on note que comme dans Equer le pourcentage de passages ne contenant aucun synonyme est très élevé : 69% pour Trec8 et 71.6% pour Trec9. Et en conséquence, le pourcentage de synonymes présents dans les passages est très faible.

11. 20 participants ont concourru à Trec8 soit 45 runs soumis ; 28 participants ont concouru à Trec9 soit 78 runs soumis

Pourcentage de synonymes WordNet présents	1.74
Pourcentage de mots de la question présents	32.52
Moyenne du nombre de mots par passage non vide	33
Moyenne du nombre de passages distincts non vides par question	87.4

Tableau 7. *Mots de la question et synonymes présents dans tous les passages de Trec8*

Pourcentage de synonymes WordNet présents	1.37
Pourcentage de mots de la question présents	36.25
Moyenne du nombre de mots par passage non vide	21.06
Moyenne du nombre de passages distincts non vides par question	125

Tableau 8. *Mots de la question et synonymes présents dans tous les passages de Trec9*

7. Conclusion

Dans les campagnes Trec comme dans la campagne Equer, les systèmes qui obtiennent les meilleurs résultats ont recours à des connaissances sémantiques ; ils utilisent WordNet, des ontologies, ou encore des bases de données spécialisées. Et de façon intuitive et légitime, on peut penser que de telles informations sont effectivement indispensables dans un système de QR tout domaine. Cependant, peu de ces systèmes fort robustes, ont évalué les gains obtenus grâce à l'utilisation de telles informations. Mesurer l'apport d'un module ou d'une base de connaissances particulière est en effet parfois complexe ou rendu même impossible par l'architecture du système. Une voie possible pour cette évaluation est celle explorée dans cet article. A défaut de savoir exactement ce qui aurait été produit si tel module avait été absent, explorer les corpus de résultats nous informe sur les connaissances utilisées, et nous donne des indications sur les améliorations indispensables à apporter.

Ainsi, cette évaluation nous a permis de détecter que l'utilisation non contrôlée de synonymes apporte peu aux performances du système, mais que la présence d'un contexte dans les multi-termes peut permettre de discriminer les synonymes et de repérer des reformulations pertinentes. La nécessité de pouvoir désambiguïser les mots-clefs des questions apparaît donc clairement si l'on veut pouvoir utiliser les synonymes. D'autres études sont en cours pour évaluer la part de la sémantique dans la tâche de QR en élargissant la taille des passages étudiés, et aussi les relations sémantiques possibles. Il est en effet légitime de penser que la signification de la question est entièrement présente dans le document réponse, puisque celui-ci doit justifier la réponse dans les campagnes d'évaluation. Le problème est de savoir sous quelle forme la retrouver, quels traitement définir et à quel moment les appliquer.

Par ailleurs, cette étude permet aussi d'évaluer le taux de bonnes réponses qu'il est possible de trouver sans apport ou presque de connaissances sémantiques, à environ

70%. C'est un chiffre déjà élevé mais non suffisant, et l'amélioration des performances devra passer par une meilleure utilisation de connaissances sémantiques.

8. Bibliographie

[AïT 02] AÏT-MOKTHAR S., CHANOD J.-P., ROUX C., « Robustness beyond shallowness : incremental deep parsing », *Journal of Natural Language Engineering*, vol. 8, n° 3-2, 2002.

[FER 02] FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., JACQUEMIN C., MONCEAUX L., ROBBA I., VILNAT A., « How NLP Can Improve Question Answering », *Knowledge Organization*, vol. 29, n° 3-4, 2002, p. 135-155.

[ITT 01] ITTYCHERIAH A., FRANZ M., ROUKOS S., « IBM's Statistical Question Answering System - TREC-10 », *Proceedings of the Tenth Text retrieval conference*, Gaithersburg, MD, 2001, NIST.

[JAC 96] JACQUEMIN C., « A symbolic and surgical acquisition of terms through variation. », *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Springer, Heidelberg, 1996, p. 425-438.

[LAV 04] LAVENUS K., GRIVOLLA J., GILLARD L., BELLOT P., « Deux pistes complémentaires pour améliorer l'appariement Question Réponse », *Workshop : Question-Réponse. Actes de Traitement Automatique de la Langue (TALN 2004)*, Fès, Maroc, 2004, p. 403-412.

[MOL 03] MOLDOVAN D., PAŞCA M., HARABAGIU S., SURDEANU M., « Performance Issues and Error Analysis in an Open-Domain Question Answering System », *ACM Transactions on Information Systems*, vol. 21, n° 2, 2003, p. 133-154.

[MOR 05] MOREAU F., SÉBILLOT P., « Contributions des techniques du traitement automatique des langues à la recherche d'information », Publication interne n° 1690, 2005, IRISA.

[YAN 02] YANG H., CHUA T.-S., « The Integration of Lexical Knowledge and External Resources for Question Answering », *Proceedings od The Eleventh Text Retrieval Conference*, Gaithersburg, Maryland USA, 2002, NIST.