Evaluation modulaire d'un système de questions-réponses sur un corpus de questions semi-spontanées

Laurianne Sitbon[*,**] — **Laurent Gillard**[*]

[*] *Laboratoire d'Informatique d'Avignon (LIA), Université d'Avignon*
339, chemin des Meinajaries
84000 Avignon, France
[**] *Laboratoire Parole et Langage (LPL), Université de Provence*
29 av. Robert Schuman
13621 Aix en Provence Cedex 1, France
{laurianne.sitbon, laurent.gillard}@univ-avignon.fr

RÉSUMÉ. *Cet article présente une évaluation séquentielle du système de questions-réponses modulaire et stochastique SQuALIA. L'évaluation se fonde sur un corpus de question semi-spontanées obtenu en faisant poser 20 questions de référence à des adultes francophones, non francophones ou dyslexiques. Les expériences montrent que ce sont les fautes d'orthographe qui ont le plus d'impact sur les modules d'analyse. En moyenne le système parvient à ne trouver des réponses qu'à 60% des questions posées, ce qui conduit à imaginer l'intégration d'un correcteur orthographique en amont des systèmes, plus de souplesse dans l'analyse, et la conservation de l'incertitude tout au long du processus en le formalisant à l'aide d'un modèle probabiliste.*

ABSTRACT. *This paper introduces the sequential evaluation of SQuALIA question answering system, a stochastic and modular question answering system. The evaluation is based on a half-spontaneously asked questions corpus. This corpus has been constructed by making french native, non native or dyslexic spellers type 20 reference questions. The results show that orthographic mistakes are the most harmful. The average good answering rate is 60% over all users. This low performance leads to new proposal such as integrating a spell checker before processing questions, propose several answer types to questions with an uncertainty degree, and keeping this uncertainty during the process, which can be defined in a probabilistic framework.*

MOTS-CLÉS : *Questions-réponses, difficultés de langage, évaluation, modèle probabiliste*

KEYWORDS: *Question-answering, language impairments, evaluation, probabilistic model*

CORIA 2007, pp. 287–301
Saint-Étienne, 28–30 mars 2007

1. Introduction

Les intérêts en recherche d'informations s'orientent depuis quelques temps vers les systèmes de questions-réponses en langage naturel. De nombreuses campagnes d'évaluation se sont intéressées à ce nouveau défi (la campagne internationale TREC [1] inclut une tâche questions-réponses (QA) [2] depuis 1999, la campagne européenne CLEF [3] intègre la tâche QA@clef où les questions sont disponibles en plusieurs langues européennes, et enfin la campagne nationale Technolangue EVALDA [4] comporte le volet EQUER), en proposant des séries de questions dont les réponses sont à trouver dans un corpus de documents journalistiques. Cependant si les questions proposées ont le mérite d'être peu ambiguës par rapport à des requêtes classiques, leur formulation en langue naturelle dans le cadre des campagnes d'évaluation n'en est pas moins trop parfaite pour coller à la réalité de l'usage de tels systèmes. L'objectif de cet article est donc de montrer comment se comporte un système face à des questions posées par des utilisateurs *tout venant*, qu'ils aient ou non des difficultés en orthographe, voire même qu'ils soient dyslexiques ou non francophones de naissance. Une analyse fine des erreurs commises par les utilisateurs permet de tracer l'impact de différents types d'erreurs (orthographiques ou grammaticales) sur les différentes parties du système. En effet les sQR fonctionnent habituellement selon une analyse modulaire : analyse de la question, recherche de documents pouvant contenir la réponse (moteur de recherche documentaire) puis analyse en profondeur des documents trouvés pour extraction de réponses.

La première partie présente la ressource principale sur laquelle nous avons basé nos expériences, le corpus de questions semi-spontanées. Elle a été réalisée dans un souci de réutilisabilité, et le processus et le contenus sont décrits. La seconde partie décrit l'analyse des différentes étapes du sQR SQuALIA pour le traitement de ces questions semi-spontanées. La troisième partie pose les bases d'une approche probabiliste plus robuste pour les systèmes de questions réponses modulaires.

2. Collecte d'un corpus de questions semi-spontanées

2.1. *et pourquoi pas spontanées ?*

Pour évaluer l'impact d'une formulation sur le sQR, il faut en premier lieu s'assurer que le système est capable de répondre à la question "si elle est posée d'une bonne façon". Si on considère les questions des campagnes d'évaluation comme des bonnes façons, on constate d'emblée que les meilleurs systèmes peinent à répondre à plus de la moitié des questions posées. Les focus de ces questions (objets des questions) étant choisis pour le corpus associé, il apparaît peu probable que si l'on recueille des ques-

1. http://trec.nist.gov
2. http://trec.nist.gov/data/qa.html
3. www.clef-campaign.org
4. http://www.elda.org/article118.html

tions posées complètement spontanément par des personnes le système soit capable d'y répondre. D'autre part, il faudrait essayer toutes les formulations possibles afin de s'assurer que l'erreur provient effectivement de l'objet de la question (auquel cas la réponse n'est pas dans le corpus), et non pas de la forme de la question.

Une manière de régler ce problème est d'identifier des questions auxquelles le système sait assurément répondre, et de faire en sorte que les utilisateurs les posent, mais avec leurs mots, et sans possibilité de les recopier (ce qui favoriserait un bonne orthographe des noms propres notamment).

Il est d'autre part important de ne pas favoriser dans l'étude certains types de questions (locatives, numératives, etc.), afin de pouvoir généraliser les observations. De même il faut veiller à ce que toutes les difficultés d'écriture soient représentées au mieux afin de ne pas biaiser les résultats. Ainsi une question qui ne contient aucun nom propre sera intuitivement plus facile à écrire qu'une question qui contient des noms rares ou étrangers. Elle contiendra probablement moins d'erreurs. C'est pourquoi les questions retenues sont également réparties selon différents degrés de difficulté des noms propres contenus.

Enfin cette représentativité oblige à un panel de personnes assez important, chacun devant poser un nombre de questions suffisant pour pouvoir distinguer ses capacités propres des problèmes plus généraaux et inhérents au système que l'on étudie, ici SQuALIA (Gillard *et al.,* 2006a).

2.2. *Protocole expérimental*

2.2.1. *Sélection des questions*

Avant de demander à des utilisateurs potentiels de taper des questions, il est nécessaire de sélectionner les questions à leur faire taper en accord avec les recommandations évoquées précédemment

Les résultats de la campagne EQueR (Ayache *et al.,* 2006) constituent un diagnostique de notre système sur 500 questions en Français. Parmi ces questions, sont sélectionnées uniquement les 122 questions pour lesquelles SQuALIA propose au moins deux bonnes réponses parmi les 5 proposées par notre système pour chaque question étaient correctes, dont 17 questions non factuelles (14 définitoires et 3 listes). Les questions définitoires étant déjà sujettes à de nombreux débats concernant l'évaluation, nous avons décidé de les écarter. Les questions booléennes étant des cas particuliers, elles n'ont pas étés retenues pour la création du nouveau corpus.

Les 20 questions sélectionnées sont listées dans le tableau 1 et se répartissent en difficultés comme suit :

– 8 questions présentant des noms propres peu fréquents

– 8 questions avec des noms propres très fréquents (Europe) ou aucun nom propre

– 2 questions contenant des noms propres étrangers et très peu fréquents.

Les 20 questions sélectionnées se répartissent en focus comme suit :

– 5 noms de personnes

– 5 nombres

– 3 dates

– 2 lieux (1 ville, 1 pays)

– 5 divers (monnaie, distance, âge, nom de journal, grade militaire)

La répartition équitable des bonnes réponses sur les questions factuelles montrée dans (Sitbon *et al.,* 2006) permet une bonne représentativité des types de questions dans le corpus. Cette répartition permet une réutilisabilité pour d'autres utilisations de ce corpus, et atteste de la généralisation des analyses et des résultats présentés ci-après.

GF222	Qui est le maire de Bastia ?
GF30	Combien de personnes souffrent d'acné en Suisse ?
GF266	Quelle est la monnaie nationale en *Hongrie* ?
GF219	A combien de kilomètres de Paris se trouve la gare de Tours ?
GF178	Comment s'appelle le président Tchétchène ?
GF6	Quel âge a l'abbé Pierre ?
GF232	Combien y a t il de chômeurs en Europe ?
GF245	Qui est le frère de la princesse Leia ?
GF17	Combien y a t il d'habitants en *Lettonie* ?
GF29	Quel grade occupe *Juan Carlos Rolon* dans la marine ?
GF273	Quand est mort *Kurt Cobain* ?
GF105	Quel est le nom du roi du *Maroc* ?
GF298	Quelle est la capitale de Terre Neuve ?
GF147	En quelle année *Hitler* est arrivé au pouvoir ?
GF176	Qui est le président d'*Aérospatiale* ?
GF99	Combien de personnes sont mortes dans des accidents de la route en 1997 ?
GF132	Où se situe *San Cristobal* de Las Casas ?
GF206	Quand a été votée la loi *Evin* ?
GF84	En combien de langues a été publié le Petit Prince ?
GF78	Quel journal publie chaque année le top 50 des personnalités ?

Tableau 1. *Questions sélectionnées dans EQueR pour créer le corpus de questions semi-spontanées, dans l'ordre où elles ont été demandées aux utilisateurs. Les mots en italique sont ceux absents du dictionnaire français de Aspell.*

2.2.2. *Récupération du corpus*

La récolte des données s'est faite à travers une application web, ce qui permet aux utilisateurs d'effectuer l'expérience dans les conditions de leur choix et dans un temps non limité, et ainsi de diminuer les facteurs liés au stress. Elle est composée d'une page d'instructions et de pages dynamiques pour la récupération des données directement dans une base de données. Pour chaque question à poser, les utilisateurs écoutent une instruction audio, qui leur explique l'objet de la question à poser. Par exemple, *on veut connaître le nom du maire de Bastia*. Les instructions ont été réalisées dans un souci de neutralité de la forme de la question finale. En effet il est difficile de ne pas influencer

l'utilisateur dans le choix des mots qu'il utilisera, comme par exemple pour *Qui est le maire de Bastia* et *Comment s'appelle le maire de Bastia ?*. Les instructions sont toutes de la même forme (*on veut connaître... on cherche à savoir...*). Dans certains cas (selon les navigateurs utilisés), on a pu enregistrer automatiquement les temps entre la dernière écoute de l'instruction et la validation de la question tapée, ainsi que le nombre de fois où chaque instruction est écoutée. Ces données ont été recueillies en cas de réutilisation du corpus dans un autre cadre, mais ne seront pas examinées ici. De même l'âge et le sexe des participants n'est pas retenu, tout en sachant qu'il s'agit uniquement d'adultes. Les participants sont classés dans 3 grandes catégories :

– 9 adultes francophones (F)

– 6 adultes non francophones (NF)

– 2 adultes dyslexiques (D)

Les adultes non francophones sont de langue natale chinoise, germanique ou hispanique et vivent en France.

2.3. *Observations préliminaires*

Les questions tapées ont été évaluées manuellement selon 5 catégories d'"erreurs" qu'elles pouvaient contenir. Les erreurs typographiques, de type absence de point d'interrogation à la fin de la question, les fautes d'accentuation (dues à un manque de connaissance ou à un clavier américain), les erreurs grammaticales ou syntaxiques (fautes d'accord ou ordre des mots), les fautes d'orthographe sur les noms propres, ou les fautes d'orthographe sur les autres mots de la question. Le tableau 2 résume les résultats de cette analyse quantitative en donnant pour chaque type de participant et chaque catégorie d'erreur le pourcentage de personnes l'ayant commise, et le pourcentage total de phrases qui la contiennent.

	?		accent		syntaxe		NP		orth	
	Npers	Nphr	Npers	Nphr	Npers	Nphr	Npers	Nphr	Npers	Nphr
N	33	13	67	15,5	67	12	100	18	44	4,5
E	50	31	83	19	100	36	100	39	100	28,5
D	50	35	100	12,5	100	25	100	30	100	17,5

Tableau 2. *Répartition des erreurs par type et population : pourcentage de personnes ayant commis l'erreur au moins une fois (Npers), et pourcentage de phrases contenant l'erreur (Nphr)*

La répartition des erreurs montre que, si les dyslexiques ou les apprenants font plus d'erreurs que les francophones non dyslexiques, des problèmes d'écriture existent aussi chez ces derniers, et donc la question de la robustesse des systèmes concerne potentiellement tout le monde. Dans l'ensemble les apprenants font plus d'erreurs que les dyslexiques. L'absence de point d'interrogation a été relevée mais n'est en

général pas un obstacle pour les systèmes. Enfin, les erreurs les plus fréquentes pour l'ensemble des sujets sont les fautes sur des noms propres.

Les erreurs de syntaxe sont essentiellement des fautes d'accord en nombre, et dans de plus rares cas des erreurs sur l'ordre des mots. D'autre part les conditions d'utilisation du système montrent que certaines erreurs comme les accents manquants peuvent se produire en fonction d'un contexte non linguistique mais technique (clavier américain par exemple, ou encore touches de clavier inopérantes).

Certaines erreurs n'ont pas été comptabilisées même si elles pourraient affecter certains systèmes. Il arrive fréquemment que les points d'interrogation soient collés au dernier mot de la question, ce qui pourrait perturber certains systèmes qui supposent un découpage en mots. Les majuscules des noms propres sont souvent ignorées (ce qui peut empêcher pour certains systèmes leur détection en tant qu'entités nommées ou noms propres). De même il arrive que les majuscules en début de question soient absentes. Ceci n'est pas une faute grave mais peut poser des problèmes si la première lettre est accentuée. Certains mots, comme par exemple *maire*, qui possède beaucoup d'homophones (*mère, mer*), sont souvent remplacés par un de ceux là, par tous les types d'individus.

3. Analyse des échecs de SQuALIA

3.1. *Les sQR modulaires*

On distingue deux grandes catégories de sQR : les sQR uniquement à bases de règles et ceux à base de méthodes stochastiques qui utilisent des règles pour l'étiquetage sémantique. Les premiers utilisent généralement des patrons de reformulation des questions afin de rechercher directement les réponses dans leur forme attendue. Dans ce cas là, la détermination de la capacité du système à répondre dépend principalement de sa capacité à reformuler, et de la présence des reformulations dans les documents. Nous nous sommes donc intéressés à la seconde catégorie de systèmes. La figure 1 illustre le fonctionnent général de ces sQR, et plus particulièrement de celui que nous utilisons, SQuALIA (Gillard *et al.,* 2006a), qui est modularisé comme la plupart des sQR aujourd'hui. Ce système ayant obtenu de bons résultats lors de la campagne internationale CLEF en 2006 (Gillard *et al.,* 2006b), son étude permettra de généraliser les résultats à l'ensemble des sQR modulaires.

Les principales étapes de traitement sont l'analyse de la question, la recherche de documents pouvant contenir la réponse (moteur de recherche documentaire) puis une analyse en profondeur des documents trouvés pour l'extraction de réponses. La plupart des sQR à base de méthodes stochastiques utilisent différents scores, notamment au niveau de l'appariement. De plus, tous utilisent une étape de focalisation à l'intérieur des documents pour cibler la ou les phrases contenant la réponse potentielle.

En pratique, après avoir étiqueté la question en type de réponse attendu (TRA) à l'aide de patrons, SQuALIA se concentre sur les documents issus d'une recherche

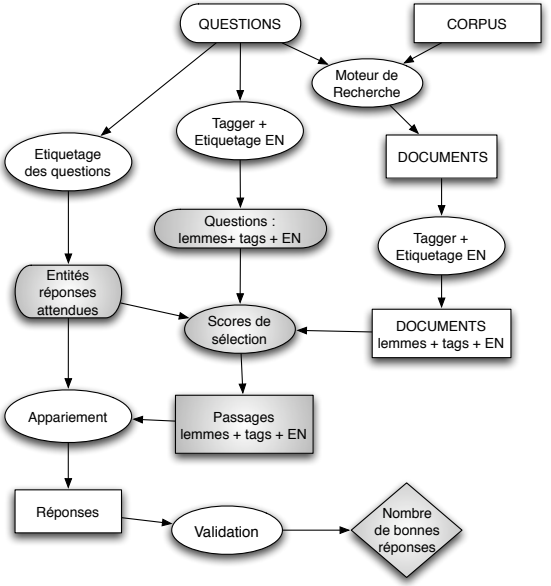

Figure 1. *Fonctionnement général des sQR.*

documentaire. Puis un score de densité est calculé pour chaque passage de ces documents, qui prend en compte une distance entre les mots et entités nommées de la question et du passage. Puis pour les passages ayant les meilleurs scores de densité, un calcul de compacité de toutes les entités nommées correspondant au type de réponse attendu est effectué. Ce calcul se fonde notamment sur la distance avec les mots de la question situés dans une fenêtre réduite autour de l'entité cible.

(Moldovan *et al.,* 2003) propose une évaluation des modules d'un sQR afin de tracer causes des erreurs qui apparaissent en fin de chaîne de traitement, pour des questions factuelles issues de la campagne TREC. De la même manière, nous allons pour chaque module tenter de déterminer les types d'erreurs qui causent des échecs pour les 20 questions auxquelles on sait que SQuALIA est capable de répondre (si elles sont posées comme dans EQueR).

3.2. *Résultats intermédiaires*

Les modules possèdent chacun des jeux d'entrées et sorties bien définis qui peuvent faire l'objet d'évaluations indépendantes, à condition d'avoir des références. Pour l'analyse en type de réponse attendues, le référentiel est celui des 20 questions

initiales vérifié manuellement. Pour les modules intermédiaires de détermination des documents puis des passages pertinents, les scores retournés sont un bon indice de réussite, comme le montre l'étude de la prédiction de la capacité d'un système à répondre à une question présentée dans (Sitbon *et al.,* 2006). Enfin, les réponses issues de l'étape finale peuvent être évaluées automatiquement à l'aide de patrons des réponses souhaitées (données de la campagne d'évaluation, puis manuellement pour celles qui ne répondent pas aux patrons).

3.2.1. *Analyse de la question*

Une analyse de la question aboutissant à un type de réponse attendue erroné, ou n'aboutissant pas du tout, implique nécessairement un échec de la réponse (Sitbon *et al.,* 2006). Cette première étape est donc primordiale. Les étiquettes posées sont celles de la hiérarchie de Sekine (Sekine *et al.,* 2002), sachant qu'on appose toutes les étiquettes d'une branche de la plus fine trouvée à la plus générale à chaque question.

L'étape d'étiquetage aboutit à 20% de questions non étiquetées et 16% sous spécifiées (étiquetage incomplet comme par exemple *nom de personne* au lieu de *président*), soit en tout à 36% d'erreurs. Le graphique de la figure 2 montre la répartition des erreurs en nombre par utilisateur, ordonnés par nombre de phrases comportant une erreur d'étiquetage. En abscisse les types d'utilisateurs (francophones, non francophones ou dyslexiques) montrent que la quantité d'erreurs produites par le module d'analyse ne peut pas être déduite du profil de l'utilisateur. D'autre part, si les étiquettes sous spécifiées n'empêchent pas le système de trouver une bonne réponse, les absences d'étiquette bloquent totalement le processus de recherche de la réponse et aboutiront à une réponse nulle.

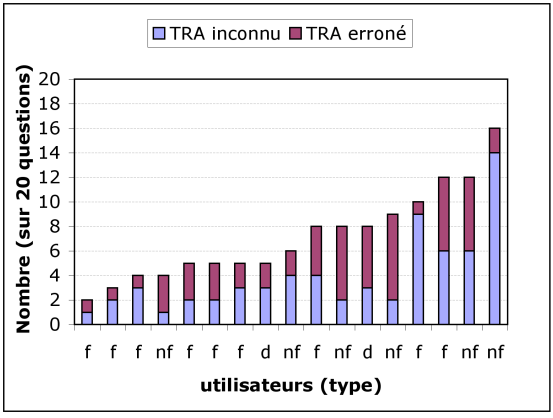

Figure 2. *Répartition des erreurs d'étiquetage (erroné ou absence d'étiquetage), pour chaque utilisateur.*

On notera que 18% des phrases mal ou non étiquetées ne comportent pas de faute. Dans ces cas là les problèmes d'étiquetage sont dus à la formulation utilisée. Pour les autres cas, la table 3 référence pour chaque type d'erreurs le pourcentage de cas où cette erreur apparaît en même temps qu'une erreur d'étiquetage. Cela ne signifie pas nécessairement une corrélation, étant donné que la plupart des phrases contiennent plusieurs erreurs. De plus les erreurs de ponctuation n'ont pas été traitées ici étant donné que le module d'analyse ne tient pas compte de la ponctuation, mais uniquement du début de la question. Les résultats montrent une plus forte implication des erreurs d'accentuation ou d'orthographe "classique" (autres) dans les phrases mal étiquetées que dans les phrases correctement étiquetées. Ce type d'erreurs peut généralement être corrigé à l'aide d'un correcteur orthographique.

accent	syntaxe	NP	autres
64,3%	46%	40,7%	66,7%

Tableau 3. *Corrélations entre erreurs et échec de l'étiquetage en type de réponse attendue : pourcentage de phrases mal étiquetées lorsque le type d'erreur y apparaît, pour chaque type d'erreur.*

Dès l'exécution de ce premier module du système, on constate une baisse de l'efficacité du système pour tous les utilisateurs : pour la plupart entre 20 et 40 % de baisse de possibilité de bonne réponse. Les causes sont variées et ne sont pas nécessairement liées à des erreurs de l'utilisateur.

3.2.2. *Recherche de documents*

Lors de la campagne EQueR, la phase de recherche documentaire était effectuée en partie par le moteur de recherche Pertimm et en partie manuellement. Les paramètres de sélection des documents pertinents n'ayant pas été explicités, il n'est pas possible de refaire cette étape avec les questions semi-spontanées. Cependant, afin de pouvoir utiliser nos résultats de EQueR en tant que référence sur les 20 questions sélectionnées, nous avons choisi de ne pas utiliser un autre moteur de recherche et d'utiliser les documents fournis lors de la campagne pour chaque question, même si en réalité on aurait des documents différents pour chaque formulation différente d'une même question (les mots clés pour la recherche étant issus de la question tapée).

La phase de recherche de documents n'est donc pas évaluée ici. Cependant le problème a été étudié dans le cadre de la recherche documentaire, notamment lors de la tâche *Confusion Track* de TREC-5 (Kantor *et al.,* 1997), où les mots des documents étaient bruités selon des schémas similaires à l'OCR. (Ruch, 2002) montre une forte dégradation du système de recherche d'informations SMART lorsque des erreurs (insertions, suppressions, substitution) sont introduites au niveau des requêtes.

3.2.3. *Sélection des passages et des réponses*

La sélection des passages se fait à l'aide d'une mesure de densité qui prend en compte les mots et les entités de la question, ainsi que le type de réponse attendu. En

réalité tous les passages contenant au moins une entité du même type que la réponse attendue sont sélectionnés quelque soit leur score, mais le score de densité sera un élément prépondérant dans la sélection des réponses. La sélection des réponses se fait à l'aide d'une mesure de compacité appliquée à toutes les entités candidates (du même type que le type réponse attendu) des passages sélectionnés. Les éléments considérés pour la compacité sont les mots de la question entourant l'entité candidate.

	densité	compacité
réponse fausse	0,075	0,512
réponse juste	0,058	0,306

Tableau 4. *Moyenne des différences entre les scores de compacité des bonnes puis des mauvaises réponses par rapport aux réponses aux questions d'origine.*

Une première manière d'analyser les échecs au niveau de ces deux modules est de comparer l'évolution des scores de densité et de compacité entre ceux des réponses aux questions d'origine, et ceux des nouvelles réponses, justes ou fausses. A cet effet, le tableau 4 contient les écarts moyens entre les scores d'origine et les scores des nouvelles réponses, classés en deux catégories (réponses justes et réponses fausses). On constate que dans tous les cas les scores des questions reformulées sont moins bons que les scores d'origine. Même si l'écart moyen pour les réponses fausses est pour les deux scores moins du double de celui pour les réponses justes, cela n'est pas suffisant pour affirmer la représentativité de ces écarts

Une autre façon de savoir si l'étape de sélection des réponses est discriminante est d'évaluer non pas les 5 mais toutes les reponses qu'on aurait propose a toutes les questions, ce qui permet d'avoir le maximum ideal des réponses auxquelles on aurait pu répondre avec des scores bien calculés. En effet, 13% des questions qui avait un type de réponse attendu connu n'auraient pas eu de réponse possible, quel que soit le calcul des scores, ce qui pose un maximum théorique de 72% de bonnes réponses sur l'ensemble des questions posées.

3.3. *Résultats finaux*

Au final, le système ne répond en moyenne qu'à 12 questions par utilisateur (soit 60% des questions posées), 10 pour les dyslexiques et les non francophones, 14 pour les francophones. le graphique de la figure 3 montre la répartition des bonnes réponses du système pour chaque utilisateur repéré par sa catégorie.

La répartition des bonnes réponses par type d'erreur montre que pour 39% des phrases contenant des fautes d'accent le système est parvenu à trouver une bonne réponse, alors que c'est le cas de 59% des phrases avec des fautes de syntaxe, 56% des phrases avec des noms propres mal orthographiés, et seulement 31,25% des phrases avec des fautes d'orthographe sur les autres mots. Les fautes d'accent étant un cas particulier de faute d'orthographe, les résultats permettent de dire que la syntaxe et

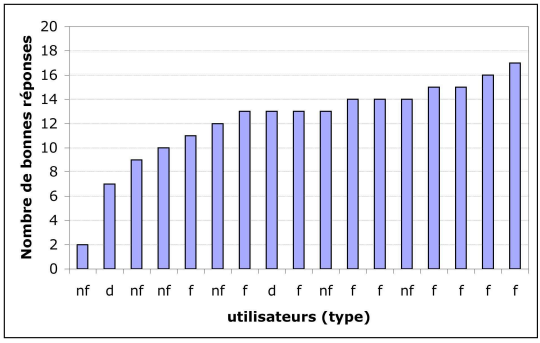

Figure 3. *Nombre de questions auxquelles le système a fourni une bonne réponse au moins parmi les 5 propositions, pour chaque utilisateur.*

l'orthographe correcte des noms propres ne sont pas les facteurs les plus déterminants pour permettre de retrouver la réponse, à condition d'avoir un moteur de recherche assez robuste pour fournir les bons documents supports.

4. Perspectives : vers un sQR plus robuste

Les modules d'analyse de la question à l'aide de règles sont fiables mais ont tendance à être incomplets. Celui de SQuALIA en est un bon exemple puisque ses performances dépassent 90% pour des questions bien écrites et dont le type est référencé dans la hiérarchie de Sekine (Sekine *et al.,* 2002), telles que les questions factuelles de EQueR et CLEF 2004, 2005 et 2006 (Gillard *et al.,* 2006a). Cependant, face à des questions spontanées, on constate une chute des performances assez significative. En effet dans beaucoup de cas le système ne propose aucune étiquette, mais lorsqu'il en propose une elle est au moins partiellement correcte. Le système pourrait être rendu plus robuste en doublant les modules d'analyse à base de règles de modules automatiques. Un tel système a été mis en place par (Gillard *et al.,* 2006b) pour la tâche multilingue avec des questions en anglais et des documents en français. Cependant de tels systèmes commettant plus d'erreurs, et pouvant parfois retourner plusieurs analyses contradictoires, l'incertitude dans ces cas là doit être prise en compte dans la suite du processus.

4.1. *Utilisation d'un correcteur orthographique pour corriger la requête*

L'ensemble des modules pourrait tirer parti d'un système de correction orthographique. Des tests ont été effectués à l'aide du correcteur orthographique Aspell [5], dont les performances sont comparables à celles des systèmes commerciaux. Les résultats montrent que si l'on considère les trois premières hypothèses pour chaque mot corrigé par l'outil, le taux d'erreurs mots moyen passe de 12,3% à 8,7%.

Cependant Aspell ne corrige pas les mots qui appartiennent au lexique, donc pas les fautes de syntaxe ou d'accords. D'autre part les noms propres correctement écrits en italique dans le tableau 1 ne sont pas dans son dictionnaire et ne peuvent donc pas être proposés en correction s'ils ont été mal orthographiés. Comme la plupart des correcteurs orthographiques fonctionnent sur la base d'un lexique fermé, il faudrait l'enrichir en amont de l'ensemble des noms propres contenus dans le corpus indexé par le sQR.

La correction grammaticale n'apparaît pas nécessaire dans la mesure où seuls les mots clés lemmatisés sont utilisés. Cependant, les correcteurs les plus performants s'appuient sur la structure syntaxique de la phrase, et dans les cas d'utilisateurs non francophones leur efficacité est diminuée. De même (James *et al.,* 2004) montre que les correcteurs orthographiques grand public sont moins efficaces pour des dyslexiques. Dans ces deux cas, la conservation de plusieurs hypothèses de correction augmente considérablement les chances de trouver le bon mot. Ainsi, il ne faut plus considérer la question comme une suite de mots mais comme un graphe d'hypothèses de mots. Un score peut être associé à chaque hypothèse en fonction de la distance entre le mot tapé et le mot proposé. Encore une fois, il s'agit de prendre en compte l'incertitude dans le processus.

L'incertitude de la compréhension des mots de la requête est un problème étudié en recherche d'informations dans le cadre de requêtes audio sur des bases de documents écrits. (Wolf *et al.,* 2002) propose une représentation des requêtes dans le modèle vectoriel qui prend en compte l'incertitude en affectant des poids à chacun des mots en accord avec le réseau de confusion produit par un système de transcription automatique. De même la recherche d'information multilingue génère des incertitudes sur les mots traduits. Cependant, la plupart des sQR multilingues ne prennent pas en compte cette incertitude.

4.2. *Approche probabiliste pour la modélisation des systèmes de questions réponses robustes*

La notion d'incertitude se corrèle inversement à la notion de probabilité de compréhension. Ainsi on peut définir un modèle probabiliste de la recherche de réponses à une question dans un corpus. Ce modèle s'inspire des modèles probabilistes pour la recherche d'information proposés par (Robertson *et al.,* 1980) et dont l'efficacité a

5. http://www.aspell.net

été attestée par (Jones *et al.,* 2000) à l'aide des pondérations d'Okapi. Les réponses recherchées sont des entités repérées. Il peut s'agir de mots ou de groupes nominaux.

Pour chaque entité e du corpus, on calcule la probabilité que ce soit la bonne réponse sachant la question posée $P(e = OK|q)$. Or on peut estimer que $e = OK$ si au moins e est du même type que le type de réponse attendu ($Te = Tq$) et si le document et le passage contenant e sont proches des mots de la question posée ($Ee = Eq$). Ces deux événements étant indépendants l'un de l'autre, on peut réécrire la probabilité que e soit la bonne réponse :

$$P(e = OK|q) = P(Te = Tq|q) \times P(Ee = Eq|q) \quad [1]$$

Si l'on considère l'ensemble T des entités de la hiérarchie de Sekine, et $T = \{T_1, ..., T_i, ...T_n\}$, alors $Te = Tq$ s'il existe un Ti tel $Te = Ti$ et $Tq = Ti$. Or Te et Tq sont indépendants donc

$$P(Te = Tq|q) = \sum_{T_i \in T} P(Te = T_i|q) \times P(Tq = T_i|q) \quad [2]$$

$P(Te = T_i|q)$ peut être estimé soit de manière binaire, en affectant 1 si le système de détection d'entités nommées a étiqueté e par T_i et 0 sinon, soit si le système de détection d'entités nommées affecte un score, en utilisant une fonction de ce score. $P(Tq = T_i|q)$ peut être estimé en fonction de la méthode qui a permis de déterminer Tq. Si c'est un patron, on peut affecter 1, si c'est un système par apprentissage tel que les arbres de décision, on peut affecter une probabilité à chaque choix possible en fonction des scores donné par le classifieur.

$P(Ee = Eq|q)$ peut être estimé à l'aide du modèle de recherche documentaire, en l'appliquant au document, puis au passage. [....] les mots $M = \{m_1...m_n\}$ de la question sont les mots d'origine plus les mots corrigés (on pourrait aussi ajouter des mots issus d'un processus d'expansion de requête). Chacun de ces mots a une probabilité d'être effectivement dans la question (ils peuvent aussi être des corrections erronées). Ainsi en appliquant cela au modèle probabiliste, et en considérant l'ensemble d_j des documents contenant e et l'ensemble p_j des passages contenant e, on a

$$P(Ee = Eq|q) = \sum_{m_i \in M} P(m_i|q) \times f(W_{m_i,d_j}, W_{m_i,p_j}) \quad [3]$$

$P(m_i|q)$ est estimé à 1 si m_i est un mot tapé par l'utilisateur et appartenant au lexique, et est fonction du score du correcteur si il s'agit d'une hypothèse de correction d'un mot tapé et hors vocabulaire (ou d'un score d'expansion si il s'agit d'une expansion de la question). W_{m_i,d_j} et W_{m_i,p_j} reflètent respectivement les scores de pertinence des documents et des passages contenant e par rapport aux mots de la question posée. La pertinence des documents peut être exprimée par la similarité des documents à la requête, et la pertinence des passage peut être estimée par les scores de compacité (à condition de ramener le passage à une fenêtre restreinte autour de l'entité considérée).

Ce modèle permet d'affecter un score de probabilité à chaque réponse possible du système, et ainsi de définir automatiquement un score de confiance sur chaque réponse proposée. On peut y ajouter des événements indépendants tels que des scores issus d'une base de connaissance, ou des possibilités de reformulation de la question. Cependant cela reste un modèle générique, et les fonctions de combinaison des différentes estimations des probabilités, ainsi que la manière d'exprimer les probabilités en fonction des scores du système restent à déterminer de manière empirique sachant un ensemble de documents, de questions et de bonnes réponses. Les différentes campagnes d'évaluation en questions-réponses apportent un tel référentiel.

5. Conclusion

Nos travaux ont permis la constitution d'un corpus de questions semi-spontanées qui sera disponible pour l'ensemble de la communauté. Ce corpus a permis de vérifier que les questions posées dans les campagnes d'évaluation ont un caractère académique qui ne permet pas complètement de corréler les performances des systèmes dans ce cadre à une utilisation dans des conditions réelles, hors de considérations sur la nature des questions posées.

Une analyse modulaire d'un système de questions réponses dont les performances générales sont à l'état de l'art a permis de montrer que si les fautes sur l'orthographe des noms communs avaient plus d'impact que les fautes d'accord ou de noms propres, elles dégradent toutes indépendamment les différentes parties du système. les fautes d'orthographe commises par des utilisateurs avaient toutes un impact, lors de différentes étapes du système. Cela mène à la conclusion que c'est chaque module du système qui doit être robuste.

Une manière de formaliser cela est d'intégrer la notion d'incertitude dans un modèle probabiliste de la recherche de réponses à une question. L'implémentation et l'évaluation de ce modèle sont les prochaines étapes de notre travail.

6. Bibliographie

Ayache C., Grau B., Vilnat A., « EQueR : the French Evaluation campaign of Question Answering system EQueR/EVALDA », *5th international Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, p. 1157-1160, May, 2006.

Gillard L., Sitbon L., Bellot P., El-Beze M., « Dernières évolutions de SQuALIA, le système de Questions/Reponses du LIA », *Traitement Automatique des Langues (TAL)*, 2006a.

Gillard L., Sitbon L., Blaudez E., Bellot P., El-Bèze M., « The LIA at QA@CLEF2006 », *Cross Language Evaluation Forum (CLEF) 2006*, Alicante, Espagne, Septembre, 2006b.

James A., Draffan E., « The accuracy of electronic spell checkers for dyslexic learners », *PATOSS bulletin*, August, 2004.

Jones K. S., Walker S., Robertson S. E., « A probabilistic model of information retrieval : development and comparative experiments », *Information Processing and Management : an International Journal*, vol. 36, n° 6, p. 779-808, Novembre, 2000.

Kantor P. B., Voorhees E., « Report on the TREC-5 confusion track », *TREC-5 Text REtrieval Conference No5,*, Gaithersburg, MD , ETATS-UNIS, p. 65-74, 1997.

Moldovan D., Pasca 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, p. 133-154, Avril, 2003.

Robertson S. E., van Rijsbergen C. J., Porter M. F., « Probabilistic models of indexing and searching », *3rd annual ACM conference on Research and development in information retrieval*, Cambridge, England, p. 35-36, 1980.

Ruch P., « Using contextual spelling correction to improve retrieval effectiveness in degraded text collections », *Proceedings of the 19th international conference on Computational Linguistics*, vol. 1, Association for Computational Linguistics, Taipei, Taiwan, p. 1-7, 2002.

Sekine S., Sudo K., Nobata C., « Extended Named Entity Hierarchy », *Proceedings of The Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, p. 1818-1824, 2002.

Sitbon L., Gillard L., Grivolla J., Bellot P., Blache P., « Vers une prédiction automatique de la difficulté d'une question en langue naturelle », *13ième conférence Traitement Automatique des Langues Naturelles (TALN)*, Louvain, Belgique, p. 337-346, 10-13 Avril, 2006.

Wolf P., Raj B., « The MERL SpokenQuery information retrieval system », *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 2, p. 317-320, Août, 2002.