

Vers une prédiction automatique de la difficulté d'une question en langue naturelle

Laurianne Sitbon ^{1,2}, Jens Grivolla ¹, Laurent Gillard ¹, Patrice Bellot ¹, Philippe Blache ²

(1) Laboratoire d'Informatique d'Avignon
339, chemin des Meinajaries - Agroparc BP 1228
84911 Avignon Cedex 9 - France
Tel : +33 (0) 4 90 84 35 09

(2) LPL-CNRS, Université de Provence
29 Avenue Robert Schumann
13621 Aix-en-Provence - France

{laurianne.sitbon, laurent.gillard, jens.grivolla, patrice.bellot}@univ-avignon.fr, pb@lpl.univ-aix.fr

Résumé Nous proposons et testons deux méthodes de prédiction de la capacité d'un système à répondre à une question factuelle. Une telle prédiction permet de déterminer si l'on doit initier un dialogue afin de préciser ou de reformuler la question posée par l'utilisateur. La première approche que nous proposons est une adaptation d'une méthode de prédiction dans le domaine de la recherche documentaire, basée soit sur des machines à vecteurs supports (SVM) soit sur des arbres de décision, avec des critères tels que le contenu des questions ou des documents, et des mesures de cohésion entre les documents ou passages de documents d'où sont extraits les réponses. L'autre approche vise à utiliser le type de réponse attendue pour décider de la capacité du système à répondre. Les deux approches ont été testées sur les données de la campagne Technolanguage EQUER des systèmes de questions-réponses en français. L'approche à base de SVM est celle qui obtient les meilleurs résultats. Elle permet de distinguer au mieux les questions faciles, celles auxquelles notre système apporte une bonne réponse, des questions difficiles, celles restées sans réponses ou auxquelles le système a répondu de manière incorrecte. A l'opposé on montre que pour notre système, le type de réponse attendue (personnes, quantités, lieux...) n'est pas un facteur déterminant pour la difficulté d'une question.

Abstract This paper presents two methods for automatically predicting the ability for a question answering system to reply a factoid question automatically. The context of this prediction is the determination of the need to initiate a dialog with the user in order to focus or reformulate the question. The first method is an adaptation of a document retrieval prediction system based on SVM and decision trees. The involved features are question or document text, and cohesion measures between documents or extracts from which the answer is extracted. The second method uses only expected answer type to predict the answer validity. Both methods have been evaluated with data from the participation of our QA engine in the Technolanguage EQUER campaign. On one hand, the SVM based method leads to the best results. It right determines which are easy questions, those our system gives the right answer, and which are hard questions, those our system gives bad or no answer. On the other hand, we show that for our system, the expected answer type (proper nouns, numbers, locations ...) is not a deterministic factor in the determination of question hardness.

Mots-clefs : Questions-réponses, prédiction de la difficulté, SVM, arbres de décision

Keywords: question-answering, difficulty prediction, SVM, decision trees

1 Introduction

Les systèmes de questions-réponses (sQR) sont au cœur des préoccupations en recherche d'information. 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 Technolanguage EVALDA ⁴ comporte le volet EQUER, auquel nous nous sommes plus particulièrement intéressés.

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. Les campagnes d'évaluation évoquées précédemment montrent que l'état des systèmes ont encore beaucoup de mal à répondre à certaines questions. Hors de toute mesure quantitative liée à des campagnes d'évaluation, la prédiction de la capacité à répondre à une question posée (qu'elle soit liée au sQR lui-même ou à l'absence de réponse dans le corpus) s'impose par un seuil d'admission de la question telle qu'elle est posée. Dans le cas où on ne peut pas répondre, il faut entamer un processus de dialogue, dans la direction adaptée, que ce soit un enrichissement (en cas d'ambiguïté sémantique) ou une correction (il vaut mieux proposer différentes alternatives que proposer une réponse à une question automatiquement corrigée), ou encore une validation de la compréhension. Il y a alors deux problèmes qui se posent dont seul le second sera traité ici : d'une part quelle doit être la nature de ce dialogue et comment en exploiter la teneur et d'autre part à partir de quel critère initier ou non le dialogue ?

L'analyse de la capacité des systèmes à apporter une information donnée est pratiquée dans le domaine de la recherche documentaire, où on cherche à affecter des scores de confiance à des résultats en fonction des requêtes posées. L'objet de cet article est d'étudier l'adaptabilité de ces méthodes à l'affectation de scores de confiance sur des réponses à des questions.

Dans une première partie nous décrivons la problématique de l'affectation de scores de confiance en recherche documentaire, et son application avec des classifieurs dans le cadre d'un sQR. La seconde partie s'attache à l'application de ce système dans le cadre de la campagne EQUER avec le sQR LIA-QA.

2 Une méthode de prédiction appliquée à un sQR

2.1 Les scores de confiance en recherche documentaire

La prédiction de la capacité à répondre à des requêtes ad-hoc est un domaine de recherche émergent, qui a fait l'objet d'un atelier ⁵ lors de la conférence internationale SIGIR en 2005 et de la tâche *robust* de la campagne TREC depuis 2003 (Voorhees, 2003). Les premiers critères de prédiction dégagés se basent sur des caractéristiques de la requête uniquement. C'est le cas de la méthode d'évaluation de la difficulté des requêtes proposée par (Loupy & Bellot,

¹<http://trec.nist.gov>

²<http://trec.nist.gov/data/qa.html>

³www.clef-campaign.org

⁴<http://www.elda.org/article118.html>

⁵<http://www.haifa.il.ibm.com/sigir05-qp/>

2000), où la fréquence de chaque mot de la requête ainsi que la fréquence combinée sont prises en compte. Les travaux de (Cronen-Townsend *et al.*, 2002) se fondent sur un calcul du taux d'ambiguïté de la requête. Dans le même esprit, (Mothe & Tanguy, 2005) ont montré que les corrélations entre des caractéristiques linguistiques de la requête et la capacité des systèmes participant à TREC 3, 5, 6 et 7 se situent uniquement au niveau de la complexité syntaxique (distance entre les termes syntaxiquement liés) et de la polysémie, écartant ainsi l'utilisation du nombre de certains types de termes (acronymes, noms propres, conjonctions, mots suffixés, ...). Ensuite l'utilisation des documents retournés par la requête a permis également de dégager de nouvelles caractéristiques pour évaluer la difficulté des requêtes. C'est le cas de (Amati *et al.*, 2004) qui étudie la répartition des termes de la requête dans les premiers documents retournés. (K.L.Kwok, 2005) utilise une régression à l'aide des SVM sur des critères appris sur les résultats de son système, qui sont la répartition moyenne de la fréquence des termes de la question, ainsi que leur quantité d'information (inverse document frequency). Les travaux de (Grivolla *et al.*, 2005) utilisent des classifieurs qui se fondent sur l'apprentissage de caractéristiques issues à la fois des questions et des documents retournés. Nous avons adapté cette dernière méthode à la prédiction de la capacité à répondre dans le cadre de la campagne EQUER, avec le sQR développé au LIA (Gillard *et al.*, 2005).

Par rapport à l'évaluation de la qualité des réponses, la campagne TREC QA s'est intéressée en 2002 (Voorhees, 2002) à la capacité des systèmes à juger de la pertinence de leurs réponses. Pour cela les participants n'avaient le droit de fournir qu'une seule réponse courte par question, mais ces réponses devaient être ordonnées par confiance décroissante. Ainsi une mesure a complété le pourcentage de bonnes réponses et les classiques rappel/précision : *Confidence Weighted Score*. Elle calcule la moyenne du taux de bonnes réponses à tous les rangs dans le classement des réponses de chaque participant. Pour répondre à cette problématique la plupart des systèmes se basent sur un consensus entre plusieurs propositions de réponses faites par différentes parties du système, notamment lors de l'utilisation de ressources externes telles que des bases de connaissances (Chu-Carroll *et al.*, 2002) ou (Bellot *et al.*, 2002). Notre système utilise également d'autres critères dans le cas de consensus, appris de manière empirique sur le comportement du système (critère sur le type de réponse attendue), ou utilisant directement des scores de recherche d'informations internes au sQR.

2.2 La prédiction par classification

Dans le système de prédiction sur lequel nous nous sommes penchés, le problème de prédire si le système est *a priori* capable de répondre à une question est vu comme un problème de classification des questions en deux classes, les questions difficiles et les questions faciles. Les algorithmes d'apprentissage utilisés pour entraîner les classifieurs sont les arbres de décision (voir par exemple (Lefèbure & Venturi, 2001) ou (Kuhn & Mori, 1995) pour leur application au traitement automatique des langues) et les machines à vecteurs supports (SVM) introduites dans (C.Burges, 1998). Ces choix sont motivés dans l'article qui présente la conception du système sur des requêtes ad-hoc de TREC (Grivolla *et al.*, 2005).

Les deux principales étapes dans la mise en place d'un tel système sont la détermination des critères de classification et le choix des données d'apprentissage.

Un certain nombre de caractéristiques de la question, sans prétraitement, ont été utilisées. Il s'agit notamment de la taille de la question. Les caractéristiques portant sur l'ambiguïté des termes, dévoilées habituellement par le nombre de synonymes, le nombre de sens connus, ou

les hyponymes, n'ont pas été utilisées dans ces premières expériences.

La deuxième caractéristique utilisée porte sur la cohésion, lexicale notamment entre les documents issus du moteurs de recherche d'où sont extraits les passages, ainsi que la cohésion entre les passages eux mêmes. Cette idée part du principe qu'une forte variabilité linguistique est corrélée avec un plus grand risque qu'une partie des documents ou passages utilisés soit non pertinents par rapport à la question, et conduisent à des réponses fausses.

La similarité cosinus a été calculée pour chaque question entre les 5, 10, 15 ou 20 premiers passages de documents ou documents complets sélectionnés par le système pour extraire les réponses. La similarité est calculée pour chaque paire de documents ou de passages en fonction des termes présents dans chacun d'eux et d'un poids qui leur est attribué pour chaque terme T_i , en fonction du nombre d'occurrences du terme T_i dans le passage ou dans le document (tf_i), et en fonction de la fréquence des lemmes dans la totalité des [D] documents du corpus de la campagne (df_i). Le poids w_{ij} d'un terme T_i dans le document D_j est donné par :

$$w_{ij} = tf_{ij} * \log|D|/df_i$$

L'autre mesure de cohésion utilisée dans le système de prédiction sur la recherche documentaire est l'entropie, qu'on peut calculer sur des ensembles de documents à l'aide d'un modèle de langage. Dans l'optique de pouvoir utiliser notre système en tant qu'aide à l'utilisateur, nous avons choisi de ne pas calculer cette caractéristique, très gourmande en calculs. C'est pour la même raison que nous avons limité le calcul des similarités moyennes entre les paires de documents aux 20 premiers document

Sur des recherches ad-hoc, les scores de similarités entre les documents et la requête sont généralement utilisés pour le classement des documents, et non pas pour une évaluation dans l'absolu de leur pertinence. Différentes transformations ont été testées sur ces scores afin d'en déterminer une corrélation avec la capacité à répondre. Il s'agit par exemple de la différence entre le score du premier document et le score du nième, ou bien une moyenne des scores des n premiers documents retournés. Pour l'adaptation de cette mesure au problème des questions-réponses, nous avons uniquement les scores de densité de chaque extrait de document par rapport à la question.

2.3 Les spécificités d'un sQR

On distingue deux grandes catégories de sQR : ceux 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 fonctionnement général de ces sQR, et plus particulièrement de celui que nous utilisons (Gillard *et al.*, 2005), qui est modularisé comme la plupart des sQR aujourd'hui. Les données sur fond gris sont celles qui ont été transmises au système de prédiction pour l'apprentissage.

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 extraction de réponses.

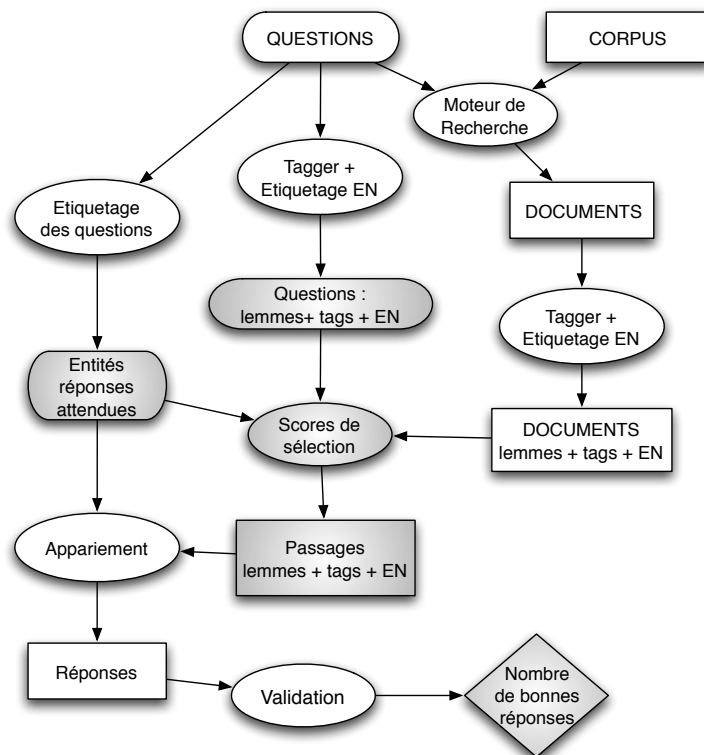


FIG. 1 – Fonctionnement général des sQR, en grisé les données utilisées par nos classifieurs

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.

3 Expériences sur les données de la campagne EQUER avec différentes méthodes de prédiction

3.1 Le moteur LIA-QA et la campagne EQUER

Lors de la campagne EQUER, pour la partie générale, les participants disposaient des 500 questions classées selon des types généraux : 30 questions pour les catégories booléennes, listes, définitoire, et 407 questions factuelles.

Etant donné que l'objet de la campagne n'était pas la recherche documentaire, les résultats d'un moteur de recherche (Pertimm) étaient fournis aux candidats. Pour chaque question les participants disposaient des 100 premiers documents trouvés. Ces résultats sont ordonnés mais les participants ne disposent pas des scores.

Pour chaque question un système pouvait retourner jusqu'à 5 réponses. Les réponses ont été validées par un ou deux juges humains.

Au niveau des classifieurs du système de prédiction, les informations utilisées sont issues d'étapes de traitement du moteur LIA-QA, décrit dans (Gillard *et al.*, 2005), et dont la figure 1 illustre le principe global.

L'étiquetage morpho-syntaxique est effectué à l'aide du Tree Tagger (Schmid, 1994). Celui en entités nommées s'appuie sur les étiquettes obtenues. Ces étiquetages sont appliqués à la fois aux questions et aux documents sélectionnés par le moteur de recherche. D'autre part des automates permettent de déterminer pour chaque question quel type de réponse est attendue (cet étiquetage est décrit plus loin). Ensuite au plus 1 000 passages de 3 phrases issues des documents sont sélectionnés, en fonction d'un score de densité prenant en compte les mots de la question, les types d'entités nommées rencontrées et le type de réponse attendu. Ensuite l'appariement est fait par le calcul d'un score de compacité, entre les étiquettes des termes contenues dans les extraits et celle attendue par la réponse.

3.2 Prédiction par arbres de classification et SVM

Les expériences ont été menées avec les arbres de classification et les SVM implémentés au sein de la boîte à outils WEKA (Ian H. Witten, 1999).

Les classifieurs ont utilisé principalement trois classes de *features* :

- les questions, ainsi que leurs lemmes filtrés ou non, les étiquettes morpho-syntaxiques et sémantiques des mots qu'elles contiennent, ainsi que le type de réponse attendue. Sur ces données sont utilisées des mesures numériques et qualitatives ;
- les documents issus de la recherche effectuée par le moteur Pertimm dans le cadre de la campagne EQUER et les passages de 3 phrases sur lesquelles s'appuie l'appariement, le tout au format lemmatisé et filtré. Les scores de similarités entre les 5, 10, 15 et 20 premiers sont calculés et utilisés comme attributs ;
- les scores de densité qui ont permis de sélectionner les passages. Des mesures de moyenne sur les 5 à 50 premiers, ou d'écart type, ou la valeur à un rang donné sont obtenues sur ces scores de densité.

Enfin les classifieurs disposaient pour chaque question du nombre de réponses courtes exactes retournées par le système, entre 0 et 5, correspondant aux deux classes, un nombre de réponses correctes supérieur à 0 pour la classe "facile", ou un nombre de réponses correctes nul pour la classe "difficile".

Les travaux en recherche documentaire sur la prédiction des questions s'évaluent souvent à partir de la distance de Kendall's tau, lorsqu'on dispose de scores liés à la prédiction. On mesure alors la distance entre le classement des réponses selon leur score de difficulté par rapport au classement selon le score de précision qu'elles ont obtenu. D'autres travaux dans un domaine où on peut déterminer si le résultat est vrai ou faux, comme les questions-réponses, utilisent pour se positionner des mesures de type CWS, qui de la même manière se fondent sur un classement de scores de prédiction. Nous ne nous sommes pas basés dans notre travail sur ce type de mesure, car nous avons cherché à déterminer de manière binaire la capacité d'un système à répondre à une question. Les résultats sont donc exprimés en pourcentage de bonne prédiction, sachant qu'une question est considérée "facile" si elle a eu au moins une réponse exacte parmi les cinq propositions.

Comme il y a un nombre relativement réduit d'exemples (de questions), l'évaluation des classifications par arbres de décision et SVM est faite par une validation croisée avec 10 plis (10-fold cross-validation). De plus, nous proposons ici les résultats sur les questions factuelles uniquement, car les autres étaient présentes en quantité trop peu représentative (30 items) pour fournir des résultats cohérents. Le tableau 1 montre les résultats des deux classifieurs, avec les différentes classes de *features* utilisées. La dernière ligne correspond à une prédiction uniquement

basée sur la distribution, qui peut donc être considérée comme la prédiction de référence.

<i>feature</i>	SVM	Arbres de Décision
toutes	68,5%	62,4%
questions	69%	64,1%
documents	53,8%	49,9%
passages	52,3%	50,1%
aucune	51,6%	

TAB. 1 – résultats de la classification automatique, en pourcentage de bonne prédiction

Les résultats montrent que l'utilisation des données relatives aux questions uniquement, quelle que soit la méthode de classification, est très efficace et suffisante. De plus la classification avec les SVM donne de très bons résultats bien supérieurs à la référence. Une des raisons possibles au succès de la classification dans notre expérience est qu'il y a dans l'apprentissage à peu près autant d'exemples positifs que d'exemples négatifs, le pourcentage de bonnes réponses obtenues dans la catégorie des questions factuelles avoisinant les 50%. On peut supposer d'autre part que l'apport très pauvre des caractéristiques reliées aux scores des similarités entre documents et passages, et de celles reliées aux scores de densité des passages, est dû au fait que ce sont les mêmes mesures utilisées par le système lui-même pour extraire les réponses, et que donc il cherche dans tous les cas à les maximiser, indépendamment du taux d'ambiguïté éventuellement amené par les mots de la question.

Les arbres de décision ont fourni des résultats moins bons que les SVM, mais en observant leur composition nous nous sommes aperçus qu'il y avait beaucoup de surapprentissage. On pourra à l'avenir optimiser cette classification en minimisant le nombre d'attributs et la taille des feuilles. La figure 2 montre un exemple d'un tel arbre, avec 8 feuilles qui prédisent si la réponse sera bonne ou mauvaise. Entre crochets est indiqué le nombre de questions correspondant effectivement à la prédiction, puis le nombre de mauvaises prédictions. Le paramétrage qui a permis d'obtenir cet arbre aboutit à 63,145% de prédiction correcte sur les questions factuelles avec toutes les classes de features disponibles en 10-fold cross-validation, et 73,7101% sur le corpus d'apprentissage. L'inconvénient majeur de ce type d'optimisations est que dans ce cas le système pourrait être trop adapté aux données de EQUER et aux résultats de LIA-QA.

3.3 Classification à partir des étiquettes de type de réponse attendue

Le type de réponse attendue a été utilisé par un certain nombre de systèmes tels que celui de (Loupy & Bellot, 2000) lorsque les systèmes ont un fonctionnement dédié. Nous avons donc tenté dans cette section de déterminer si cette indication peut être un critère, sinon déterminatif, au moins prépondérant dans la détermination de la difficulté des questions.

Le composant d'étiquetage en types de réponses attendues de LIA-QA dédié à l'appariement est un étiquetage hiérarchique des questions. La hiérarchie utilisée a été inspirée par celle proposée par (Sekine *et al.*, 2002), dont elle est un sous-ensemble. Le choix de ce sous-ensemble, qui comprend 100 étiquettes, a été fait selon une observation de la fréquence des questions associées à une entrée de cette hiérarchie lors des précédentes campagnes d'évaluation QR de CLEF (2003 et 2004). Concrètement, cette phase d'étiquetage se déroule après une première étape d'uniformisation, à base de règles et de lexiques, permettant de réduire les différentes variantes à une même écriture et ainsi de diminuer le nombre de règles d'étiquetages à 172.

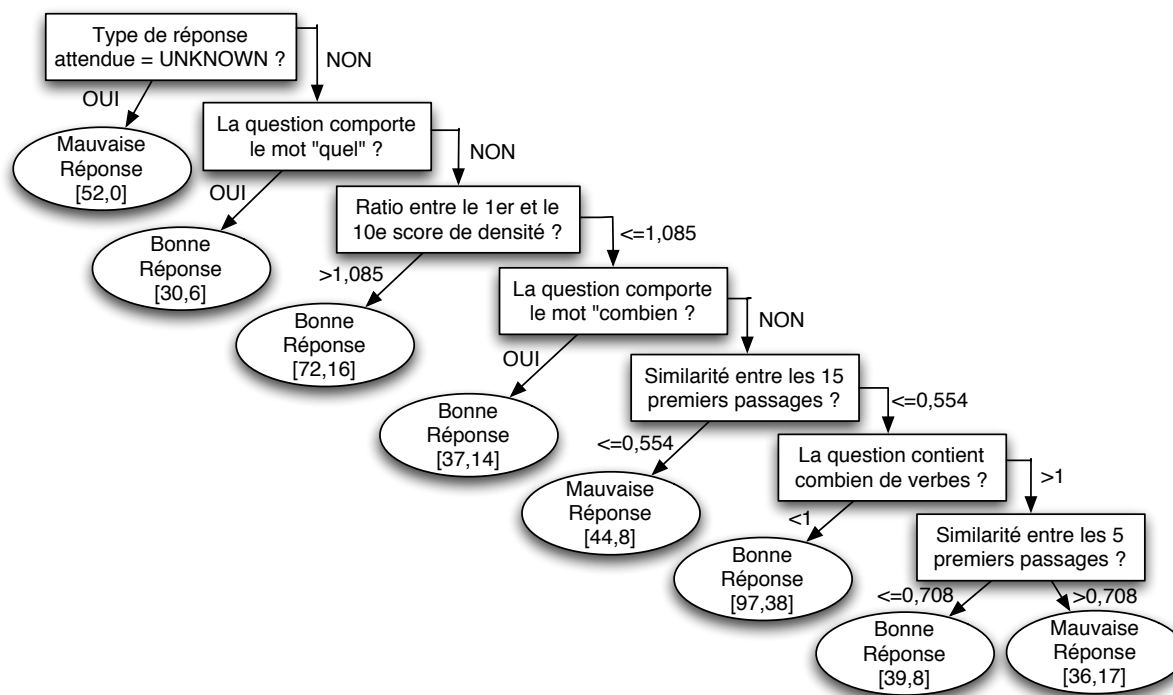


FIG. 2 – Un exemple d'arbre de décision optimisé. Les valeurs indiquées dans les feuilles sont respectivement le nombre de questions correctement classées par ce critère, et le nombre de mauvaises prédictions.

Afin de pouvoir utiliser nos résultats sur la classification à partir des étiquettes, nous avons fait une vérification manuelle de l'étiquetage en type de réponses attendues, dont voici l'analyse. Sur les 407 questions factuelles il y en a eu 52 non étiquetées dont :

- 13 étaient du type *procédé*, introduites par l'interrogatif *comment*
- 8 étaient du type *procédé* exprimées différemment (par quel procédé, de quelle façon, ...)
- 7 concernaient des événements
- 3 étaient des "pourquoi"
- 2 étaient floues en termes d'attente de réponse : "qu'en est il de ... ?" ou "que devient ... ?"

Les 19 autres questions se répartissent en 7 questions qui auraient dû être étiquetées par le système, et 12 dont les types n'étaient pas attendus même si ils se trouvent dans la hiérarchie des entités nommées de (Sekine *et al.*, 2002). Ce sont des entités de type maladie, nourriture, animal, etc.

Parmi les 355 étiquettes de questions proposées, seulement 10 étaient erronées. Parmi elles 3 étaient des numérations d'une mauvaise cible, et 3 n'auraient pas dû être étiquetées car elles sont de type *procédé*.

Le tableau 2 recense pour certains types de réponses attendues le pourcentage de bonnes réponses fournies par le LIA-QA. On constate que, à part pour certains types très peu représentés, comme les films, les causes de décès ou les fonctions par exemple, la plupart des types conduisent à l'obtention d'une bonne réponse dans environ 57% des cas. Cela montre qu'on ne peut pas utiliser uniquement le type de réponse attendue pour déterminer a priori si on sera capable de répondre à la question, et donc qu'on ne peut pas déterminer la difficulté d'une question, pour notre moteur en tous cas, en fonction de l'objet de la question. Cela dit, même si cette caractéristique des questions n'est pas suffisante comme seul prédicteur, elle était utilisée

TYPE	N	OK	% R
Lieu	76	43	56,6%
nationalite	12	9	75%
ville	10	8	80%
Nombres	81	48	59,3%
argent	6	4	66,7%
longueur	10	6	60%
employés	4	1	25%
population	7	6	85,7%
personnes	10	7	70%
age	7	3	42,9%
Fonction	8	1	12,5%

TYPE	N	OK	% R
Personne	102	62	60,8%
president	9	6	66,7%
écrivain	7	5	71,4%
Organisation	18	8	44,4%
parti politique	8	7	87,5%
entreprise	2	1	50%
Date	41	24	58,5%
Journal	4	3	75%
Cause de décès	4	3	75%
Film	3	0	0%

TAB. 2 – résultats des questions factuelles par type de réponse attendue : N est le nombre de questions étiquetées par le type, OK est le nombre de questions pour lesquelles il y a eu au moins une bonne réponse fournie, et %R le pourcentage de ces questions.

par les classifieurs dans les expériences précédentes, et elle peut tout à fait avoir contribué à une bonne classification. Il est à noter que pour toutes les questions où l'étiquette n'a pas été déterminée, le système n'a pas fourni de réponse. C'est donc le seul critère que nous avons pu dégager à l'aide de l'étiquetage en type de réponse attendue.

4 Conclusion et perspectives

Les deux approches que nous avons mises en place pour la prédiction de la capacité d'un système à fournir une bonne réponse à une question factuelle ont montré que l'élément le plus significatif à prendre en compte était la question elle-même, et non pas les documents utilisés pour trouver la réponse ou encore le type de réponse attendue. Cela est plutôt une bonne chose car cela montre que sans faire activer le moteur de questions-réponses on peut obtenir une très bonne indication sur la nécessité de la préciser ou de la reformuler. De plus les résultats obtenus avec les SVM sont très encourageants, et ils pourront probablement encore être améliorés par l'intégration de nouvelles ressources.

En effet nous avons très peu adapté la classification au problème particulier des questions-réponses, afin de vérifier dans un premier temps si les solutions pouvaient coïncider avec celles des systèmes de recherche d'informations *ad hoc*. L'intégration du taux d'ambiguïté dans les critères de prédiction à l'aide d'un réseau de collocations ou de le réseau de connaissances sémantique EuroWordNet devra être la prochaine étape de cette étude. En effet les questions contiennent généralement peu de mots, ce qui amène un plus grand risque d'ambiguïté des termes.

La difficulté des questions est un problème qui a été soulevé essentiellement du point de vue de la classification par types de réponses attendues. L'autre aspect qui peut rendre une question difficile est la nécessité d'un traitement de la négation ou de la voix passive (Lavenus & Lapalme, 2002). L'intégration de ce type de critères syntaxiques devra à terme être prise en compte. En effet même si ces problèmes là ne se rencontrent pas spécifiquement dans la campagne EQUER, ils correspondent à un besoin dans l'utilisation réelle de ces systèmes.

D'autre part les cas d'utilisation réelle amènent à s'interroger sur le cas des erreurs d'orthographe, notamment sur les noms propres (Hongrie → Ongrie), ce qui en fait des mots absents

du corpus et ne permet pas de retrouver les réponses sans traitement spécifique. Des erreurs sur les autres mots de la requête peuvent également avoir une influence importante sur la capacité à répondre. On pourrait alors imaginer un score orthographique qui selon le type d'erreurs pourraient prédire la capacité du système à passer outre.

Références

- AMATI G., CARPINETO C. & ROMANO G. (2004). Query difficulty, robustness and selective application of query expansion. In *Actes de ECIR'04*, Sunerland, UK.
- BELLOT P., CRESTAN E., EL-BÈZE M., GILLARD L. & LOUPY C. D. (2002). Coupling named entity recognition, vector-space model and knowledge bases for trec-11 question answering track. In *Actes de TREC 11*.
- C.BURGES C. J. (1998). A tutorial on support vector machines for pattern recognition. **2**(2), 121–167.
- CHU-CARROLL J., PRAGER J., WELTY C., CZUBA K. & FERRUCCI D. (2002). A multi-strategy and multi-source approach to question answering. In *Actes de TREC 11*.
- CRONEN-TOWNSEND S., ZHOU Y. & CROFT W. B. (2002). Predicting query performance. In *Actes de SIGIR'02*, p. 299–306.
- GILLARD L., BELLOT P. & EL-BÈZE M. (2005). Le lia à equer (campagne technolanguage des systèmes questions-réponses). In *Actes de TALN'05*, Dourdan, France.
- GRIVOLLA J., JOURLIN P. & MORI R. D. (2005). Automatic classification of queries by expected retrieval performance. In *Actes de SIGIR'05*, Salvador, Brazil.
- IAN H. WITTEN E. F. (1999). *Data Mining : Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.
- K.L.KWOK (2005). An attempt to identify weakest and strongest queries. In *Actes de SIGIR'05*, Salvador, Brazil.
- KUHN R. & MORI R. D. (1995). The application of semantic classification trees to natural language understanding. **17**(5), 449–460.
- LAVENUS K. & LAPALME G. (2002). Evaluation des systèmes de question réponse. aspects méthodologiques. **43**(3), 181–208.
- LEFÈBURE R. & VENTURI G. (2001). *Data mining, Eyrolles*.
- LOUPY C. D. & BELLOT P. (2000). Evaluation of document retrieval systems and query difficulty. In *Actes de Using Evaluation within HLT Programs : Results and trends*, p. 31–38, Athènes, Grèce.
- MOTHE J. & TANGUY L. (2005). Linguistic features to predict query difficulty - a case study on previous trec campaigns. In *Actes de SIGIR'05*, Salvador, Brazil.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees.
- SEKINE S., SUDO K. & NOBATA C. (2002). Extended named entity hierarchy. In *Actes de LREC 2002*, Las Palmas, Canary Islands, Spain.
- VOORHEES E. M. (2002). Overview of the trec 2002 question answering track. In *Actes de TREC 11*.
- VOORHEES E. M. (2003). Overview of the trec 2003 robust retrieval track. In *Actes de TREC 12*.