Influence de mesures de densité pour la recherche de passages et l'extraction de réponses dans un système de questions-réponses

Laurent Gillard, Patrice Bellot, Marc El-Bèze

Laboratoire d'Informatique d'Avignon (LIA) – Université d'Avignon
339 ch. des Meinajaries,
BP 1228
F-84911 Avignon Cedex 9, France
{laurent.gillard, patrice.bellot, marc.elbeze}@univ-avignon.fr

RÉSUMÉ. *Dans cet article, nous comparons différentes méthodes de filtrage et d'extraction d'une réponse candidate dans le cadre d'un système de questions-réponses. Ces expériences sont effectuées sur un sous-ensemble du corpus de la campagne Technolangue-EQueR, première campagne francophone de questions-réponses utilisant des questions et un corpus en français. Nous évaluons la méthode que nous avions retenue lors de notre participation à cette campagne. Celle-ci est basée sur une densité et une compacité des mots de la question dans le contexte d'une réponse candidate, elle est présentée et comparée à deux autres approches : l'une utilisant un décompte des mots communs, l'autre une similarité de type Cosine. Cela nous permet, également, d'envisager l'influence de la recherche de passage sur le module final d'extraction d'une réponse qui utilise cette compacité.*

ABSTRACT. *In this paper, we evaluate some answer candidates extraction algorithms for question answering. These experiments have been performed on a subset of the Technolangue-EQueR, first French question answering campaign using French corpus and questions. During our participation to EQueR, we used density and "compacity" scores, calculated using the words appearing inside a question, to filter answer candidates. These scores are described and then compared to two other techniques: one using common word counts and the other one, a Cosine similarity. Also, we study the influence of our passage retrieval algorithm and its influence on our final extraction module.*

MOTS-CLÉS : *Système de questions réponses, campagne EQuER, similarités questions / passage, similarités questions / réponses candidates, compacité, densité.*

KEYWORDS: *French Question answering, EQueR evaluation campaign, questions / passages similarities, questions / answer candidates similarities, density.*

1. Introduction

Les systèmes de Questions-Réponses (sQR) peuvent être définis comme des systèmes de recherche d'information spécifiques. En effet, la requête à traiter se présente sous la forme d'une question formulée en langue naturelle, et le résultat de la recherche doit être une (ou plusieurs) réponse concise fournie dans le cadre d'un contexte validé par un document. Ainsi, pour répondre à une question, un sQR effectue des traitements qui apparaissent comme autant de filtrages. Successivement, ils sélectionnent dans le volumineux corpus initial, un sous-ensemble de documents susceptibles de contenir une réponse, puis un ensemble de passages, et enfin un ensemble de réponses candidates dans lequel la (les) meilleure réponse est choisie. Dans cet article, nous nous intéressons plus particulièrement à ces deux derniers points : le filtrage des passages et l'extraction d'une réponse candidate.

La sélection des meilleurs passages, dans les systèmes à l'état de l'art, fait intervenir un filtrage utilisant un calcul de densité des mots de la question dans le voisinage des réponses candidates. C'est notamment le cas dans le système que nous avions développé pour notre participation (Gillard *et al.*, 2005) à la campagne EQueR, *É*valuation en *Que*stions-*R*éponses (QR), première campagne d'évaluation QR sur le français, et décrite dans (Ayache *et al.*, 2005). En outre, en plus d'utiliser une densité pour la sélection des passages, nous utilisons également une notion similaire (que nous appelons « compacité ») pour l'extraction de la réponse. L'objet principal de cet article est l'évaluation de l'influence de ces calculs de densité et une comparaison avec deux autres approches, une utilisant un décompte des mots communs et, une autre, une similarité Cosine avec une pondération TF.IDF.

La suite de ce papier est organisée comme suit : la section 2 présente des travaux connexes concernant la sélection des passages et l'extraction des réponses mais commence par un bref rappel de l'architecture d'un sQR conventionnel ; la section 3 est consacrée à la présentation des mesures de densité étudiées, la section 4 à la comparaison des différentes approches envisagées et à la présentation de résultats.

2. Travaux connexes

Un sQR peut être décrit schématiquement comme un enchaînement séquentiel de tâches caractéristiques, où chacune contribue à produire l'entrée de l'étape suivante :

– La première a pour objet la question et pour objectif de la typer en lui associant un (ou plusieurs) type de ce qui va être recherché (que nous appelons Entité Réponse attendue, ERa). Elle s'accompagne aussi de la construction d'une requête de Recherche d'Information (RI) à destination de l'étape suivante,

– ensuite, une recherche de document utilisant un système de RI conforme à l'état de l'art (tels que MG ou Lucène) permet de cibler les traitements suivants (plus complexes et longs) sur un sous-ensemble plus pertinent que l'intégralité du corpus,

– la troisième étape correspond à un autre filtrage, cette fois à un ensemble de passages susceptibles de contenir une instance des ERa,

– enfin, l'ultime étape est la sélection des meilleures réponses accompagnées de leur contexte de réalisation.

À coté de cela, un traitement supplémentaire a lieu le plus en amont possible, généralement après la RI, et permet de baliser ce qui sera envisagé comme autant d'hypothèses de réponses ou Entités Réponses candidates (ERc). Il s'agit d'un étiquetage en Entités Nommées (EN), le plus fin possible en QR, et d'une correspondance entre ces EN (potentiellement des ERc) et les types d'ERa.

Les deux dernières étapes sont l'objet de nos expériences (sections 3 et 4), nous étudions ci-dessous quelques travaux connexes. Ainsi, les systèmes concourant aux campagnes d'évaluation TREC emploient pour la plupart des techniques de recherche de passages afin de réduire la taille des données à explorer par le module d'extraction de réponse. L'approche la plus simple consiste à sélectionner les passages qui ont le plus de mots communs avec la question (Light *et al.*, 2001). D'autres systèmes, tels le nôtre, tiennent compte de la distance qui séparent, dans les passages, les mots de la question (Ittycheriah *et al.*, 2001) : plus l'éloignement de ces mots est grand, moins il y a de chances qu'un lien sémantique fort les unisse (au contraire de ceux qui participent à une même « idée »). Dans le cas de questions longues, une piste à explorer consisterait à mettre en rapport la densité des mots dans un passage avec la densité de ces mêmes mots dans la question : une forte proximité de certains mots dans un passage peut être un handicap si les occurrences de ces mêmes mots, dans la question, sont éloignées. Autrement dit, ça n'est peut être pas tant le fait que les mots de la question soient proches dans un passage qui font que celui-ci est à retenir mais le fait qu'ils sont « distribués » pareillement (Chauché *et al.*, 2003). Une densité allant dans ce sens pourrait s'exprimer par un rapport des distances entre les mots apparaissant à la fois dans le passage et la question.

Une alternative consiste à relever les co-occurrences communes entre la question et un passage. Plusieurs auteurs ont proposé d'utiliser des modèles de langage afin de tenir compte des relations existant entre les mots d'une question et favoriser les passages dans lesquels des relations identiques apparaissent (Gao *et al.*, 2004). Mais, outre la nécessité de disposer d'un corpus d'apprentissage étendu – ce qui est peu envisageable avec des questions ouvertes –, ces approches sont limitées du fait qu'elles se situent uniquement au niveau lexical et occultent la structure de la phrase. Par exemple, dans la question : « *Quand a été construite la maison d'arrêt de Fleury-Mérogis ?* », l'interrogation concerne la maison d'arrêt *située à* Fleury-Mérogis et non pas celle d'une autre ville. Cette dépendance fonctionnelle doit être retrouvée dans les passages candidats : lorsque le sujet est lié à un complément du nom par exemple, les mots correspondant dans le passage candidat doivent être proches, mais ils doivent aussi être reliés par une relation grammaticale identique ou très proche. (Cui *et al.*, 2005) ont formulé ce constat et montré en outre l'apport significatif de cette approche par rapport aux méthodes uniquement lexicales.

Enfin, bien qu'un travail similaire (à celui que nous présentons dans ce papier), ait déjà été effectué sur un sous corpus particulier des campagnes TREC-QA par (Tellex et al. 2003), nous souhaitons contrôler la faisabilité d'un filtrage et d'une sélection des meilleures réponses grâce à des scores basés sur des densités pour le

français et valider ainsi les intuitions qui avaient dictées nos choix lors de notre participation à EQueR. Nous renvoyons à (Tellex et al. 2003) pour plus de détails sur un état de l'art concernant les techniques de sélection des réponses candidates employées dans les campagnes anglo-saxonnes TREC-QA.

Par ailleurs, ce travail nous permet également d'évaluer les performances des phases intermédiaires de notre sQR. En effet, un des artefacts qui accompagne les campagnes d'évaluation en QR est que les performances mesurées sont finalement des performances globales et en bout de chaîne. Il est d'ailleurs intéressant de constater que la dernière campagne TREC-QA 2005 s'interroge sur les liens entre recherche des passages et extraction des réponses (au travers d'une tâche « document ranking » où les systèmes doivent produire la liste ordonnée des documents qu'ils utilisent pour répondre aux questions). Ainsi, notre démarche se situe dans le cadre d'une post-évaluation de la campagne EQueR telle celle menée par (Perret, 2005).

4. Densité et Compacité

Répondre à une question peut être envisagée comme un appariement entre :

– une question dont les termes sont autant d'indices permettant un « filtrage » et une sélection du contexte, cela afin d'aboutir à la meilleure réponse envisageable,

– et une réponse candidate, c'est-à-dire une bribe de texte cohérente, extraite d'un document, le plus souvent sous la forme d'une Entité Nommée, qui plus est d'une nature correspondant au type de la question (Entité Réponse attendue, ERa).

À partir de ce constat, nous proposons de définir une notion de densité permettant de choisir un extrait en fonction de son contexte commun avec une question. Il est possible d'envisager cette « compacité » à différents niveaux : du document vers le passage, pour extraire les meilleurs sites informatifs, et du passage vers la réponse, pour extraire de ces passages la réponse la plus crédible.

4.1. Densité pour le filtrage des passages

Pour calculer la densité qui permet de réduire l'espace de recherche des documents vers des passages, un ensemble d' « objets caractéristiques » (o_i, fig. 2) est extrait de la question (Q) afin de constituer une sorte de requête, il est constitué :

– des lemmes des mots à l'exception de ceux des mots outils,

– des types d'Entités Nommées (EN) présentes,

– et du (des) type(s) d'Entité Réponse attendue lorsque celle-ci est une EN.

Ensuite, pour chacun des « objets caractéristiques » rencontrés o_i, (X dans la fig. 3), à l'intérieur de chaque document, une distance moyenne $\mu(o_i)$, évaluée en « nombre d'objets », est calculée entre l'occurrence courante o_i et celles des autres objets de la requête (β, Y et δ dans la fig. 3), ou de leur plus proche occurrence en cas de présences multiples (β dans la fig. 3). Ces calculs de distance ne considèrent pas les limites de phrases, mais bien toutes les occurrences dans le document.

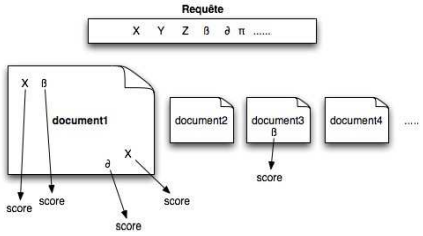


Figure 2. *Objets caractéristiques et calcul du score de densité pour chacun d'entre eux*

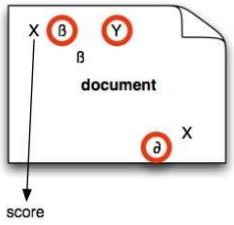


Figure 3. *densité de o_i=X calculée en fonction de β, Y et δ*

Cette distance moyenne est utilisée dans un calcul du score de densité effectué pour chacun des « objets caractéristiques » (fig. 2) d'un document *D* :

$$score(o_i, D) = \frac{\log\left[\mu(o_i) + \left(\text{card}\{\bigcup_{o_i \in Q} o_i\} - \text{card}\{\bigcup_{o_i \in Q \cap D} o_i\}\right) \times pénalité\right]}{\text{card}\{\bigcup_{o_i \in Q} o_i\}} \quad [1]$$

Ainsi, ce score de densité permet d'identifier la zone d'un passage qui présente le plus de similitudes avec une question. En effet, pour chaque position candidate d'une phrase (une position candidate est une position où un objet caractéristique de la question est présent), la densité est estimée en fonction de la distance qui sépare cette position candidate des autres objets caractéristiques, du nombre de ces objets dans la question, et enfin, du nombre d'objets communs entre une question et le document en cours. La pénalité, fixée empiriquement, a pour rôle de favoriser plus ou moins une forte proximité de quelques objets communs avec la requête par rapport à une proximité plus faible d'un plus grand nombre d'objets communs. Au contraire, lorsque tous les objets de la requête sont trouvés, la pénalité ne doit pas intervenir. Enfin, le score d'une phrase est établi comme étant le plus élevé des scores des « objets caractéristiques » qu'elle contient. Chaque phrase est ensuite étendue à un « passage » constitué de la phrase précédente à la phrase suivante (lorsqu'elles existent). Cela afin d'essayer de compenser une éventuelle perte de contexte sur des phrases courtes ou utilisant des référents trans-phrase (aucune résolution d'anaphore n'est employée dans les expériences présentées plus loin[1]). Le fait d'avoir défini un passage comme un bloc de 3 phrases est empirique (même si cela correspond à une intuition acceptée en QR ; une autre possibilité aurait été de faire varier la taille de ce passage en fonction de l'éloignement des objets caractéristiques). Les meilleurs passages suivant ce score et au plus 1000, pour l'ensemble des documents trouvés sont proposés à l'entrée de l'étape suivante.

[1] Néanmoins une expérience sommaire nous a montré que le remplacement des pronoms personnels anaphoriques « il/elle » par le 1er nom qui les précède permettait un léger gain.

4.2. Compacité pour la sélection de la réponse

Les différentes Entités Réponses candidates (ERc) à confronter à ce stade sont présentes dans des passages jugés informatifs au sens qu'ils contiennent un contexte « dense » susceptible de permettre le choix de la meilleure parmi ces hypothèses. Afin d'y parvenir, nous employons un critère inspiré du CWS (*« Confidence Weighted Score »*, défini dans Voorhees, 2002), soit le critère d'évaluation employé lors de TREC-11. L'idée est de considérer chaque occurrence d'une ERc comme un point zéro d'un repère, et la présence des mots de la question autour de cette ERc, comme des réponses correctes, les mots apparaissant dans son voisinage et non présents dans la question sont alors considérés comme des réponses incorrectes. Un critère fondé sur un calcul reproduisant celui de la précision moyenne n'est pas satisfaisant : si une question contient *n* mots et qu'un seul est présent dans un passage, il aura tendance à attribuer un score plus important à un passage contenant un seul mot à côté de l' ERc qu'à un passage contenant outre ce mot à la même position d'autres mots de la question sur des positions plus éloignées. Aussi, pour compenser cela, il suffit de modifier légèrement ce critère de précision moyenne pour y introduire une once de la notion de rappel en divisant par le nombre de mots différents de la question. En revanche, un des défauts de ce critère est qu'il favorise une question courte par rapport à une question longue. Cependant, ce défaut n'a pas d'influence dans nos expériences puisque ce critère n'est utilisé que pour ordonner les réponses à une même question entre-elles (et non pas l'ensemble des réponses à toutes les questions, ce qui était demandé lors de TREC-11). Notre critère peut être décrit comme une compacité moyenne normalisée des réalisations des mots de la question au voisinage d'une occurrence d'une ERc. Ce qui est recherché est idéalement le plus compact et complet des sacs de mots (Luhn, 1958) provenant de la question autour des frontières gauche et droite d'une ERc.

Soit *MQ* l'ensemble des mots non vides de la question, soit *X* l'un de ces mots, soit ERc_i une entité réponse candidate ; cette compacité se définit comme :

$$compacité(ERc_i) = \frac{\sum_{X \in MQ} p_{X,ERc_i}}{|MQ|} \qquad [2]$$

Où $p_{X, ERci}$ correspond à la précision à l'intérieur d'une fenêtre centrée sur l'Entité Réponse candidate ERc_i pour les mots non vides de la question à l'intérieur d'un rayon *R* fixé par l'occurrence la plus proche X_p du mot *X* (cf. algorithme 1 et [3]).

pour chaque $X \in MQ$:
  soit X_p l'occurrence de *X* la plus proche de ERc_i
  $R=distance(X_p, ERc_i)$
  $W=\{Y \mid distance(Y, ERc_i) \leq R \text{ et } Y \in MQ\}$

$$p_{X,ERC_i} = \frac{|W|}{2R+1} \qquad [3]$$

Algorithme 1. *Calcul de la précision*

Il est à noter que l'unité de base est ici le mot, aussi les éventuelles collocations ne sont pas considérées dans leur globalité. De même, l'ordre des mots n'est pas pris en compte (malgré notre intuition que, à part pour les inversions verbes/sujets dues à la formulation des questions, cet ordre est important). En revanche, cela autorise une variabilité des expressions comme l'insertion ou l'omission d'un mot. En complément, il pourrait être intéressant de considérer la longueur des différentes ERc (afin de faire intervenir une notion de concision de l'hypothèse), et celles des questions (une question longue propose plus de contexte et pourrait être plus facile). Enfin, dans l'implémentation actuelle, le rayon de la fenêtre de compacité est fixé par l'occurrence la plus proche du mot X_p, il serait intéressant d'étudier à quel point cette proximité peut masquer une possible meilleure contribution d'une occurrence plus lointaine, éventuellement en dehors du passage…

5. Résultats Expérimentaux

5.1. Cadre d'évaluation

Les données des expériences décrites ci-après proviennent de la campagne d'*É*valuation en *Que*stions-*R*éponses (EQueR) du projet EVALDA/Technolangue (http://www.technolangue.net/article61.html et Ayache, *et al.*, 2005).

Les questions considérées sont un sous-ensemble des questions factuelles de la tâche « généraliste » : soit celles avec une réponse dans la collection de documents (sinon la réponse est « *NIL* ») ; qui n'ont pas été retirées de l'évaluation (suite à des erreurs d'orthographe) ; et qui sont correctement étiquetées par notre analyseur de question (à base de règles ; dans le cas contraire, de part l'enchaînement séquentiel, l'extraction ne peut aboutir) ; soit au total 335 questions sur les 395.

Plutôt que de considérer l'intégralité du corpus, les organisateurs d'EQueR proposaient, pour chaque question, la liste des 100 meilleurs documents retournés par le moteur de recherche Pertimm. Les expériences décrites ici utilisent ces documents, exception faite de ceux provenant du Sénat (leur mise en forme, taille et disproportion posent problèmes aux autres composants ; lors de la campagne la plupart des participants les ont d'ailleurs écartés pour ces mêmes raisons)[2].

Avant de poursuivre, quelques notions doivent être éclaircies : comment sont définies une *réponse courte* ou une *réponse passage correcte* ? Au sens des campagnes d'évaluation et pour nos expériences, pour qu'une réponse courte soit jugée *correcte*, il est nécessaire qu'elle soit *exacte*, pour cela, elle doit contenir seulement l'information nécessaire pour répondre à la question (ni ajout, ni omission sur les chaînes de caractères) ; ET être *supportée*, c'est-à-dire que le document dont elle est extraite permette de l'établir. Une réponse de type passage est par essence

[2] Les documents restants proviennent des 3 sources suivantes : « Le Monde » et de « Le Monde Diplomatique », des années 1992 à 2000, ainsi que les dépêches de presse de l'agence télégraphique Suisse « Schweitzeriche Depeschenagentur » ; l'ensemble est disponible auprès de ELDA (http://www.elda.org/).

moins précise et correspond à un bloc d'au plus 250 octets ; elle est jugée correcte si elle contient une réponse courte correcte. Enfin, par opposition, une réponse peut être : *inexacte*, lorsque la chaîne contient trop (ou pas assez) d'informations ; *non supportée*, dans le cas où son document d'origine ne la justifie pas et enfin *incorrecte* si aucune des conditions n'est remplie.

À ce jour, il n'existe pas de référence officielle permettant une évaluation sur le jeu EQueR. D'ailleurs, à notre connaissance, la seule référence en QR est en anglais (Lin *et al.*, 2005). Cependant, depuis les campagnes TREC-QA, et hors de celles-ci, les sQR sont couramment évalués par le biais de motifs d'expressions régulières (et de leurs identifiants de documents respectifs) dérivés des réponses connues comme correctes et supportées. Il est généralement effectué deux décomptes des réponses : l'un dit « strict » correspond à celui des réponses correctes et supportées ; l'autre dit « tolérant » correspond aux réponses reconnues par l'un des motifs (sans l'appui d'un document support ; cela pour pallier le manque d'exhaustivité des références qui empêche notamment les calculs de rappel et précision). L'ensemble des deux est considéré comme un intervalle borné qui permet de situer les performances d'un sQR. Afin d'évaluer les méthodes ci-après, nous avons construit une telle référence (disponible sur demande) depuis l'ensemble des réponses des participants de la campagne EQueR et nous l'utilisons pour les expériences décrites plus bas.

5.4. Résultats sur le filtrage de passage

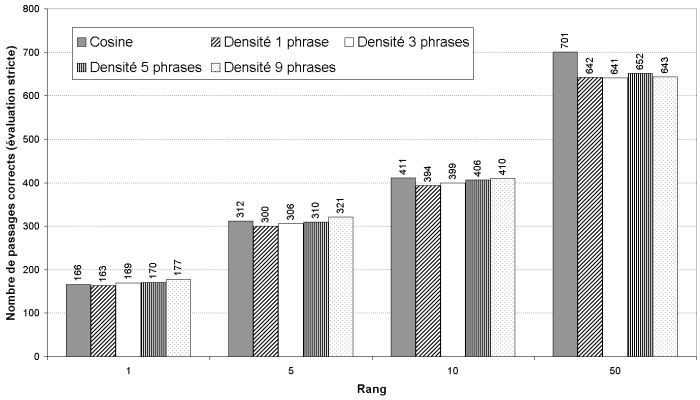
Le tableau 1 présente une comparaison entre le filtrage de passages que nous proposons et un filtrage utilisant une similarité Cosine (avec une pondération TF.IDF calculée sur l'ensemble des documents en entrée). Les valeurs présentées sont le nombre de passages (de 3 phrases) contenant une réponse correcte en fonction de leur rang. Ainsi, pour un filtrage basé sur notre densité et une évaluation stricte, 169 passages comprennent une réponse correcte au premier rang et 306 sur les 5 premiers. Il apparaît de ce tableau que la densité se comporte aussi bien, voire mieux, que Cosine pour réduire l'espace de recherche des réponses correctes.

	Cosine – évaluation		Densité – évaluation	
Rang	Stricte	Tolérante	Stricte	Tolérante
1	166	205	169	240
5	312	419	306	507
10	411	623	399	770
50	701	1634	641	2025

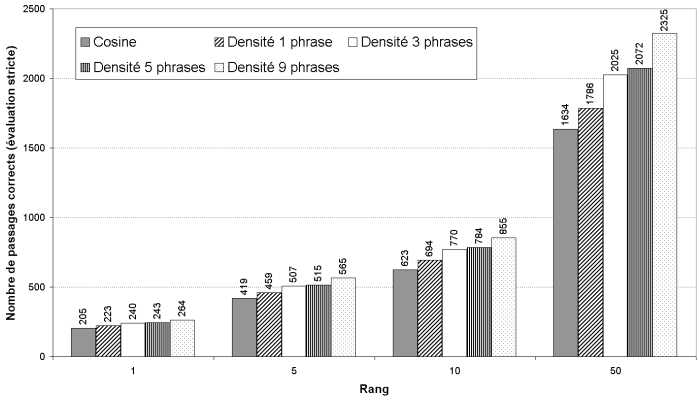
Tableau 1. *Nombre de passages (de 3 phrases) avec une réponse correcte*

Le graphique 1 reprend ces mêmes résultats pour une évaluation stricte et le graphique 2 pour une évaluation tolérante. Également, dans ces graphiques la taille des passages filtrés varie : de gauche à droite, les histogrammes correspondent à un

filtrage depuis Cosine avec extension à un passage de 3 phrases (soit la colonne « Cosine » du tableau 1) ; depuis notre densité et une extension à des blocs de 1 phrase, 3 phrases (soit comme les autres expériences décrites dans cet article et les valeurs de la colonne « Densité » du tableau 1), 5 phrases et 9 phrases.



Graphique 1. *Nombre de passages avec une réponse correcte, évaluation stricte*



Graphique 2. *Nombre de passage avec une réponse correcte, évaluation tolérante*

La lecture de ces graphiques nous permet de constater que dans le cas d'une évaluation tolérante, et conformément aux attentes, plus la taille de la fenêtre constituée par le passage s'agrandit (1, 3, 5 puis 9 phrases) plus le nombre de réponses correctes augmente. Il faut également ne pas perdre de vue que le nombre de réponses qui s'avérerait être des faux positifs devrait pareillement croître.

Concernant une évaluation stricte, en revanche on peut constater que cette croissance n'apparaît que dans le cas de la métrique basée sur Cosine, cela peut s'expliquer par deux raisons : la première, viendrait d'une effective supériorité d'un

filtrage de passage avec une similarité Cosine. La seconde, également soulignée dans les travaux de (Lin, 2005) provient de la manière dont est construite la référence utilisée pour nos évaluations. En effet, elle est déduite des résultats retournés par l'ensemble des systèmes participants ; si ces systèmes, ou même le meilleur, utilisent des similarités apparentées, leur contribution à la référence entre dans une spirale qui favorise les résultats provenant d'un système utilisant une similarité équivalente.

Nous pensons que cette dernière hypothèse est celle à retenir. En effet, suite à une évaluation manuelle dans un contexte identique à celui d'EQueR, nous avons constaté que les évaluations automatiques strictes avec notre référence sous-estimaient les résultats réels. Ainsi, la difficile post-évaluation des systèmes de questions-réponses reste posée : il est illusoire de dresser une référence exhaustive des réponses à des questions telles celles envisagées, sauf à restreindre ces questions à une part congrue, ce qui serait tout aussi critiquable. Il s'agit là d'une question ouverte soulevée par (Lin, 2005) à laquelle nous n'apportons pas d'autre réponse qu'en proposant une évaluation stricte et une autre tolérante. En, outre, bien qu'ayant participé à la campagne EQueR avec un système classé second sur les réponses courtes (et donc à l'état de l'art), notre contribution à la référence a été moindre en raison de la présence d'erreurs (corrigées depuis) ; et qui peuvent expliquer une sous estimation sur une évaluation stricte. Malgré ces avertissements nous pensons que les tendances à retenir de l'ensemble de nos expériences restent applicables.

5.5 Impact de la compacité sur la sélection des réponses

Afin d'évaluer l'impact de la compacité sur les performances de notre sQR, nous avons effectué les combinaisons suivantes dans un cadre similaire à celui de la campagne EQueR, où 5 propositions de réponses courtes étaient possibles par question. Les résultats, présentés dans le tableau 2, proposent le décompte des réponses correctes (réponses supportées pour l'évaluation stricte, et correspondance simple pour la tolérante) pour chaque combinaison. En outre, afin de disposer d'un ensemble de résultats de base, nous avons envisagé l'utilisation d'une pondération Cosine pour filtrer les passages et une sélection des réponses depuis le type d'ERa et du nombre de mots communs entre une question et le passage contenant l'ERa.

Filtrage du passage par	Sélection des réponses par	Évaluation			
		Stricte		Tolérante	
Cosine	Mots communs	91	36,5%	104	40,2%
	Compacité	158	63,5%	174	67,2%
	Cosine & Compacité	145	58,2%	166	64,1%
Densité	Mots communs	81	33,6%	94	37,6%
	Compacité	161	66,8%	177	70,8%
	Densité & Compacité	166	**68,9%**	179	**71,6%**

Tableau 2. *Sélection des réponses : nombre de réponses correctes*

Nous avons déterminé les maxima atteignables par notre système (utilisés pour le calcul des pourcentages). Ils sont, pour un filtrage des passages par une similarité Cosine, de 249 réponses courtes correctes dans le cadre d'une évaluation stricte et de 259 pour une évaluation tolérante ; pour notre densité, ils sont respectivement de 241 et 250. En effet, bien qu'une réponse soit présente dans un passage, si le sQR n'a pas la connaissance qui lui permet de l'extraire, généralement sous la forme d'un balisage correspondant à une Entité Nommée, cette réponse ne peut être proposée. De même, si, en amont, le type d'ERa n'est pas apparié (erreur d'étiquetage ou étiquette inconnue) avec ce type d'Entité Nommée, le système n'a même pas la connaissance qu'il peut retourner cette information pour la question considérée.

Le tableau 2, permet de voir que densité et compacité semblent être complémentaires, bien qu'elles reposent sur des approches similaires, puisque l'enchaînement séquentiel des deux permet l'un des meilleurs résultats, et mieux une combinaison linéaire du score de densité et de la compacité (lignes « Densité & Compacité ») permet le meilleur résultat quelle que soit l'évaluation. Enfin, il est notable que la compacité apporte un net gain quel que soit le filtrage préalable, ce qui conforte nos attentes.

Une nuance doit cependant être présentée ici. Bien que ces résultats soient encourageants, l'une des questions sous-jacente qui peut apparaître provient du « fossé » existant entre le vocabulaire utilisée dans la question et celui du contexte de la réponse : comment envisager le calcul d'une compacité dans le cas où aucun mot en commun n'existe ? Plusieurs solutions pourraient être envisagées, comme l'utilisation de synonymes provenant d'une ressource sémantique comme WordNet.

6. Conclusion et perspectives

Nous avons comparé sur un corpus en français de questions-réponses, deux approches pour filtrer au mieux les passages contenant une réponse candidate : nous avons ainsi pu vérifier la bonne tenue d'un score défini à partir d'une densité des mots de la question par rapport à une approche de référence en Recherche d'Information utilisant une similarité de type Cosine.

De même nous avons pu vérifier notre hypothèse qu'une compacité des mots de la question dans le voisinage d'une Entité Réponse candidate permettait d'extraire les meilleures d'entre elles, cela par rapport à d'autres métriques. Il reste à explorer à quel point une telle compacité pourrait se substituer à l'ensemble des trois phases de filtrage actuellement présentes et permettre ainsi de passer directement du corpus à l'ensemble des meilleures réponses. Une autre critique qui peut être faite sur ces densités est qu'elles ne prennent pas en compte l'ordre des mots, il faudrait pallier cela : en prenant en compte la position relative des mots entre eux en plus de leur proximité et de leur fréquence, et/ou en considérant plus spécifiquement des ensembles de mots (et par conséquent leur structure telle celle sous jacente à l'intérieur d'entités nommées). Par exemple, nous envisageons de calculer plusieurs compacités sur des n-uplets successifs (voire avec des synonymes) afin de capturer ces phénomènes de collocations/positions et d'en extraire une compacité globale par

une combinaison linéaire (dont les coefficients seraient fonction d'une « pertinence » de ces n-uplets ou même l'objet d'un apprentissage).

Enfin, les résultats établis sur des corpus de questions réponses en anglais ont été reproduits sur le français. Il est ainsi possible de constater que la compacité permet de préserver une indépendance vis-à-vis de la langue comme nous le recherchions.

7. Bibliographie

Ayache C., Choukri K., Grau B., *Rapport de la Campagne EVALDA/EQueR Evaluation en Questions-Réponses*, 2005. http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf

Chauché J., Violaine P., Jaillet S., Teisseire M. « Classification automatique de textes à partir de leur analyse syntaxico-sémantique », *Actes de TALN 2003*, Batz-sur-Mer, France, 11-14 juin 2003.

Cui, H., Sun, R., Li, K., Kan, M., and Chua, T. "Question answering passage retrieval using dependency relations.", *Actes de "the 28th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)"* Salvador, Brésil, 15-19 août 2005, p. 400-407.

Gao J., Nie J.-Y., Wu G., Cao G., "Dependency language model for information retrieval", *Actes de "the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR 2004)"*, Sheffield, Royaume-Uni, 25-29 juillet 2004, p. 170-177.

Gillard L., Bellot P., El-Bèze M., « Le LIA à EQueR », *Atelier de la campagne Technolangue-EQueR, Actes de TALN-Recital 2005*, Dourdan, France, 6-10 juin 2005, Tome 2, p. 81-84.

Ittycheriah A., Franz M., and Roukos S., "IBM's statistical question answering system", *Actes de "the 10th Text REtrieval Conference"*, Gaithersburg, Maryland, USA, 13-16 novembre 2001, p. 258-264.

Light M., Mann G. S, Riloff E., Breck E., "Analyses for elucidating current question answering technology", *Journal of Natural Language Engineering, Special Issue on Question Answering*, Vol. 7, No. 4., 2001.

Lin J., Katz B., "Building a Reusable Test Collection for Question Answering.", *Journal of the American Society for Information Science and Technology*, 2005, sous presse.

Lin J. "Evaluation of Resources for Question Answering Evaluation", *Actes de "the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)"*, Salvador, Brésil, 15-19 août 2005, p. 392-399.

Luhn H.P., "The Automatic Creation of Literature Astracts", *IBM Journal of Research and Development,* Volume 2, Issue 2, avril 1958, p. 159-165.

Perret L., « Extraction automatique d'information : génération de résumé et question-réponse », Thèse de doctorat, Université de Neuchâtel, 2005.

Tellex S., Katz B., Lin J., Marton G., Fernandes A., "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering", *Actes de "the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)"*, Toronto, Canada, 28 juillet – 1er août 2003, p.41-47.

Voorhees E.M., "Overview of the TREC 2002 Question Answering Track", *Actes de "the 11th Text REtrieval Conference"*, Gaithersburg, Maryland, USA, 19-22 novembre 2002.