# Quelles combinaisons de scores et de critères numériques pour un système de Questions/Réponses ?

Laurent Gillard[(1,2)], Patrice Bellot[(2)], Marc El-Bèze[(2)]

[(1)]CEA, LIST, Laboratoire d'ingénierie de la connaissance multi-média multilingue,
18 route du Panorama, BP6, FONTENAY AUX ROSES, F- 92265 France
prénom.nom @ cea.fr

[(2)]Laboratoire d'Informatique d'Avignon (LIA)
Université d'Avignon et des Pays de Vaucluse
F-84911 Avignon Cedex 9 (France)
prénom.nom @ univ-avignon.fr

**Résumé** Dans cet article, nous présentons une discussion sur la combinaison de différents scores et critères numériques pour la sélection finale d'une réponse dans la partie en charge des questions factuelles du système de Questions/Réponses développé au LIA. Ces scores et critères numériques sont dérivés de ceux obtenus en sortie de deux composants cruciaux pour notre système : celui de sélection des passages susceptibles de contenir une réponse et celui d'extraction et de sélection d'une réponse. Ils sont étudiés au regard de leur expressivité. Des comparaisons sont faites avec des approches de sélection de passages mettant en œuvre des scores conventionnels en recherche d'information. Parallèlement, l'influence de la taille des contextes (en nombre de phrases) est évaluée. Cela permet de mettre en évidence que le choix de passages constitués de *trois* phrases autour d'une réponse candidate, avec une sélection des réponses basée sur une combinaison entre un score de passage de type *Lucene* ou *Cosine* et d'un score de *compacité* apparaît comme un compromis intéressant.

**Abstract** This article discusses combinations of scores for selecting the best answer in a factual question answering system. Two major components of our QA system: (i) relevant passage selection, and (ii) answer extraction, produce a variety of scores. Here we study the expressivity of these scores, comparing our passage density score (i) to more conventional ranking techniques in information retrieval. In addition, we study varying the length (in number of sentences) of context retained in the relevant passages. We find that a three sentences window, and a mixing of *Lucene* or *Cosine* ranking with our *compactness* score (ii) provides the best results.

**Mots-clés :** Système de Questions/Réponses, compacité, densité, combinaison de scores.
**Keywords:** Question Answering, compactness, density, combination of scores.

## 1 Introduction

Les systèmes de Questions/Réponses (QR) se proposent d'extraire *LA* réponse à une question formulée en langage naturel depuis un ensemble de documents. Dans ce travail, nous nous intéressons aux systèmes de Questions/Réponses (sQR) capables de répondre à des questions

factuelles, c'est-à-dire à celles dont la réponse est un court énoncé, et plus spécifiquement l'expression d'une information sémantique précise et concise telle que le *nom d'une personne*, *d'un lieu*, la *date d'un événement*, une *valeur numérique*, *etc.* La réponse à produire est obtenue par extraction depuis un document plutôt que par synthèse. Enfin, le cadre expérimental typique pour les résultats obtenus ci-après est semblable à celui proposé par différentes campagnes d'évaluation en QR comme le sous-volet *Question Answering* de la campagne *TREC* (Voorhees, Harman, 2005) ou encore la campagne *Évaluation en Questions-Réponses* (EQueR) (Ayache *et al.*, 2005).

Le système de Questions/Réponses développé pour ce travail au LIA possède une architecture relativement générique (explicitée en 1.1). Il fait intervenir un enchaînement séquentiel de traitements pour aboutir à une réponse finale. Ces traitements sont à percevoir comme autant de processus de filtrage pour limiter le contexte d'une recherche : d'abord du corpus vers un ensemble de documents, puis de cet ensemble vers des passages, et des passages vers une réponse. Il est à noter que ce système repose uniquement sur une approche fondée sur des scores et ne nécessite que peu ou pas de connaissances syntaxiques. Concrètement, nous avons défini deux scores pour la tâche de QR : le premier score, dit de densité, permet de sélectionner des passages susceptibles de contenir une réponse intéressante pour une question; le second score, la compacité, permet, dans un passage, de sélectionner la meilleure des réponses. Ces deux scores sont calculés relativement à une réponse candidate préalablement repérée dans un document ou dans un passage, celle-ci devant être d'un type compatible avec celui attendu exprimé par la question. Ces deux scores sont utilisés en séquence : la densité intervient dans un module situé en amont de celui qui emploie la compacité.

Dans cet article, nous nous intéressons plus spécifiquement à une combinaison de ces scores de densité et de compacité. Notre objectif est d'étudier s'il est possible de faire ressortir des caractères complémentaires, *a priori* inconnus, qui permettent d'améliorer les performances en bout de chaîne et par conséquent ayant un impact sur la sélection d'une réponse, notamment par rapport à un emploi en séquence de ces scores. En outre, afin d'étudier cette notion de complémentarité de manière plus exhaustive, d'autres recherches de passages mettant en œuvre des scores classiques en recherche d'information (*Cosine*, *Cosine+Okapi* et score *Lucene*) sont également examinées et mises en concurrence. Nous vérifions qu'une combinaison de ces scores est préférable à un score obtenu sans combinaison. De plus dans le but d'élargir l'intérêt de ce travail, différentes tailles de passages[1] (de *une*, *trois*, *cinq*, ou *neuf* phrases autour d'une réponse potentiellement intéressante) sont envisagées. Par ailleurs, comme nous exploitons des scores, c'est tout naturellement vers une combinaison numérique que nous nous sommes dirigés. Enfin, d'un point de vue applicatif, il est à noter que si les deux combinaisons proposées ont bien été employées lors de nos deux dernières participations avec le LIA à des campagnes d'évaluations des systèmes de Questions/Réponses, les résultats présentés sont ceux d'expériences faites hors campagnes, sur les questions factuelles d'EQueR, et ont été de ce fait évalués au regard d'une référence tirée d'EQueR.

## 1.1 Architecture générique d'un sQR

Un système de Questions/Réponses peut être schématiquement décrit comme un enchaînement de différents modules correspondant à trois étapes principales : une analyse (à percevoir comme une *« compréhension »*) de la question, un traitement des documents et

[1] Il s'agit de la dimension des passages produits en sortie du module de sélection des passages. Lesquels fixent ensuite le contexte d'extraction pour une réponse précise à l'étape aval correspondante. C'est bien la présence (ou l'absence), en fin de chaîne, d'une réponse précise correcte qui est évaluée dans nos expériences.

enfin une extraction d'une ou de plusieurs réponses. Concernant, l'étape intermédiaire de traitement des documents, elle est très souvent décomposée en une recherche documentaire classique suivie d'une exploitation des documents sélectionnés à l'issue de cette recherche dans le but de localiser des passages susceptibles de contenir des réponses dans la perspective d'une extraction finale. La figure 1 présente cette architecture schématique à trois, et plus généralement quatre composants ainsi que leurs principales interactions. Chaque étape produit en sortie des informations utilisées en entrée de l'étape suivante (voire en entrée de plusieurs de celles situées en aval). Par ailleurs, pour quelques systèmes plus complexes, il manque à ce schéma d'éventuels flux supplémentaires générés lors d'allers-retours et/ou rétroactions entre les composants (relaxation de contraintes, *etc.*).

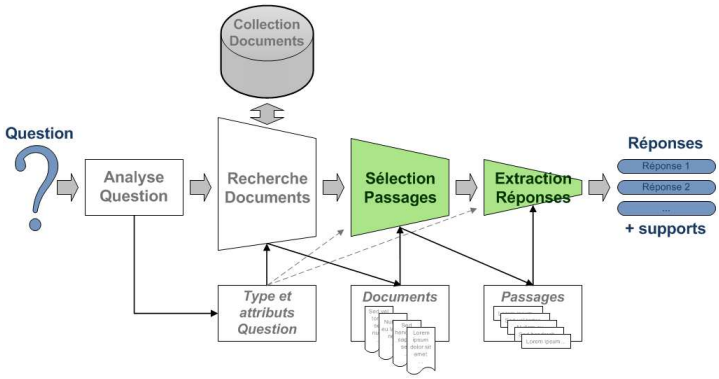



*Figure 1. Architecture générique d'un système de Questions/Réponses*

À titre d'illustration, voici un exemple du fonctionnement de cette architecture générique pour la question *« En quelle année est né Nelson Mandela ? »* :

- L'analyse de la question identifie la nature de l'information qui va être recherchée (une étiquette d'entité ou un type de réponse attendue) (Ferret *et al.*, 2002), soit ici, l'*année* associée à une *date*, et plus spécifiquement celle d'une *date de naissance*. Elle construit également une requête à destination d'un système de recherche documentaire comme par exemple, la requête constituée par les seuls lemmes *« naître »*, *« Nelson »*, et *« Mandela »* ;
- le système de recherche documentaire propose une liste d'identifiants de documents du corpus contenant tous ces lemmes, soit les documents *« ATS.940426.0081 »*, *« LEMONDE94-000873-19940509 »*, *« LEMONDE94-003506-19940429 »*, *« LEMONDE95-000054 »*, *« LEMONDE95-012364 »*. Cette liste pourrait être ordonnée suivant une similarité prenant en compte la fréquence des lemmes ;
- le composant en charge du traitement des documents effectue un balisage des entités de type *date*. Il découpe ces documents en blocs ou passages, effectue une sélection des passages intéressants (Tellex *et al.*, 2003) et élimine, par exemple, ceux qui ne contiennent ni une entité *date*, ni les lemmes *« naître »*, *« Nelson »*, et *« Mandela »* ;
- enfin, l'extraction d'une réponse consiste à choisir la « meilleure » réponse à proposer comme résultat, cela en accord avec l'étiquette sémantique associée lors de la première étape (pour cet exemple, la *date de naissance* ou son générique *date*). Cette sélection de la, ou des meilleures réponses pourrait être faite grâce à l'utilisation de patrons morphosyntaxiques ou grâce à un calcul de proximité des différents termes associés aux lemmes intéressants de la question dans les passages. Et, en définitive, le *« 1918 »* de *« Nelson Mandela est né le [$_{DATE}$18 juillet 1918] dans un village xhosa du Transkeï. »* (du document *« ATS.940426.0081 »*) apparaît ainsi comme une meilleure réponse que le *« 1912 »* rencontré dans le passage : *« Nelson Mandela _ soixante-quinze ans _ n'était pas né quand le Congrès national africain (ANC) vit le jour ([$_{DATE}$1912]). »* (*« LEMONDE94-003506-19940429 »*).

## 2 Quelle combinaison de scores, quels critères ?

Dans le système de QR envisagé, chacune des deux étapes, sélection de passages ou sélection d'une réponse, est associée à un score. Ce dernier est une valeur numérique réelle qui permet d'ordonner l'ensemble des possibles, le meilleur des scores correspond au meilleur des cas.

**Sélection de passages.** Ce score de passage est dépendant de la technique mise en œuvre pour le filtrage. Ainsi, il peut s'agir de notre score de densité (Gillard *et al.*, 2005), ou bien de scores de similarité plus conventionnels[2], que nous envisageons à titre de comparaison, tels que *Cosine*, d'une combinaison entre *Cosine* et *Okapi* (Wilkinson *et al.*, 1996), ou de celui calculé par *Lucene*. Lorsque le score considéré est notre score de densité, l'une de ses caractéristiques essentielles est d'être un score de proximité des objets de la question (les mots qui la constituent et les entités nommées qu'elle comprend) à l'échelle du document, tout en considérant, et même en englobant, une réponse candidate repérée et marquée comme susceptible d'être en adéquation avec le type attendu par la question. Dans les cas d'une similarité vectorielle, l'expressivité du score repose principalement sur des concepts de fréquence et de rareté, notamment au travers des pondérations *TF.IDF (« Term Frequency×Inverse Document Frequency »)*. De plus, *Okapi* fait également intervenir un critère de taille moyenne des passages ; et, le score *Lucene*, un ratio, appelé couverture, entre les mots de la question et ceux de ces mots présents dans le passage.

**Sélection et extraction d'une réponse.** Notre score de compacité positionnelle (Gillard *et al.*, 2007), fait appel à des notions de proximité des mots (Luhn, 1958) de la question dans un passage. Ces mots doivent en outre être au plus près de la réponse candidate (l'idée est là encore une notion de la plus grande *« densité »* possible) compatible avec le type attendu par la question mais cette fois dans une localité restreinte à un fragment d'un passage. Une autre possibilité est d'envisager un décompte des mots communs entre la question et le passage extrait, complété par une stratégie telle que celle employée lors de la participation du LIA à TREC-11 : celle-ci consistait à proposer comme réponse la première réponse compatible avec le type attendu. L'expressivité attendue pour ces différents scores peut être schématisée dans un tableau tel que le tableau 1. Bien que nous ne l'ayons pas expérimenté, il est également possible de présenter l'usage de patrons morphosyntaxiques qui correspond à une autre grande famille de type d'extraction pour une réponse dans des systèmes de Questions/Réponses.

Le **tableau 1** présente l'expressivité de ces différentes méthodes. Ainsi, la fréquence correspond à une notion habituellement associée à la fréquence d'un mot dans un document, ou plutôt, dans le cas qui nous intéresse ici, dans un passage. Il s'agit d'une notion associée aux pondérations intervenant dans des calculs de similarité. La rareté, que l'on associe aisément à l'inverse d'une fréquence au sein d'un ensemble de documents n'intervient pas dans nos calculs de compacité positionnelle, dans ceux de densité ou pour des patrons morphosyntaxiques. L'ordre des mots, est un point faible dans l'ensemble des scores envisagés ici, mais il serait capturé par des patrons morphosyntaxiques. Il est à noter qu'il est tout à fait envisageable d'effectuer des calculs de compacité sur des unités plus grandes qu'un mot et ainsi de faire apparaître une telle notion d'ordre en considérant, par exemple, des bigrammes de mots ou toutes combinaisons de n-grammes[3]. Cependant, dans le cas d'une

---

[2] Un système de recherche booléen ne permet pas différencier et d'ordonner suffisamment les passages trouvés.

[3] Nous avons commencé à emprunter cette voie dans le cadre d'une utilisation d'un score de compacité pour du résumé automatique, mais sans noter un impact significatif. Cependant, selon nous, il s'agit d'une voie à poursuivre et d'autres expériences sont à mener.

question, il existe une difficulté supplémentaire puisqu'il faut être capable de retrouver depuis une forme interrogative, l'ordre de sa forme affirmative. Enfin, si l'ordre est effectivement l'un des points forts de l'expressivité associée à des patrons morphosyntaxiques il faut remarquer que le coût d'acquisition de ces patrons n'est pas minime (même lorsque des techniques automatiques sont employées).

La connaissance d'un type de réponse attendu est une information typiquement associée à un composant d'un système de Questions/Réponses, aussi n'apparaît-elle pas pour des scores originellement issus de la recherche documentaire. De même, la proximité des mots est une notion qui nous a paru particulièrement importante dans le cadre des sQR, et que nous avons souhaité incorporer en priorité dans notre travail sur les scores. Cette proximité intervient également dans le cas de patrons morphosyntaxiques. Il est possible de noter que dans le cas des patrons, cette proximité est particulièrement rigide, au contraire de celle envisagée dans nos calculs de compacité et de densité, ce qui peut à la fois constituer un inconvénient ou un avantage selon le passage. En effet, un patron qui sélectionne une expression le fait avec une plus *« grande force »*, mais aussi avec un *« plus grand »* risque de passer à coté, parfois à cause de différences très faibles. Enfin, la proximité dans le cas de la densité est une proximité étendue à la globalité du document, alors que, pour la compacité, il s'agit d'une proximité locale au passage. De notre point de vue les deux ont une importance. Il est plus facile de trouver une réponse au sein d'un passage lorsque la formulation de ce passage est proche de celle de la question (*i.e.* emploie les mots de la question). Cependant, il est des cas, dans lesquels au moins un des mots de la question, présent dans le document, n'est plus répété par la suite et n'apparaît plus dans la fenêtre d'un passage, alors que sa présence mérite d'être prise en compte. Un exemple typique est le cas d'une question comportant une restriction sur un pays particulier, notamment dans un style de rédaction journalistique : après un premier centrage thématique sur ce pays, il est probable que son nom ne soit plus répété dans le reste du document. Enfin, la présence des mots de la question apparaît comme un ingrédient évident de tous les scores envisagés. Ce n'est pas le cas de la notion de couverture, qui n'intervient que dans nos scores, et dans celui calculé par *Lucene*. Concernant les patrons morphosyntaxiques, ils capturent cette couverture d'une façon détournée. D'autres critères ont été explorés comme la variabilité des mots, au travers d'une racinisation ou d'une lemmatisation.

| | ***Sur les mots de la question (Q)*** | | | | | | ***Type de réponse attendu*** |
|---|---|---|---|---|---|---|---|
| | ***Fréquence*** | ***Rareté*** | ***Ordre*** | ***Proximité*** | ***Présence*** | ***Couverture*** | |
| ***Mots communs*** | | | | | × | | |
| ***Cosine*** | × | × | | | × | | |
| ***Cosine+Okapi*** | × | × | | | × | | |
| ***Lucene*** | × | × | | | × | × | |
| ***Densité*** | | | | ×[1] | × | × | × |
| ***Compacité positionnelle*** | | | | ×[2] | × | × | × |
| ***Patrons morphosyntaxiques*** | | | × | ×[2] | × | × | × |

*Tableau 1. Tableau croisé des expressivités de différentes méthodes de sélection d'un passage ou d'une réponse en fonction de critères intéressants à capturer. (La proximité peut être envisagée globalement[2] à un document ou localement[1] à un passage).*

***Remarque.*** Afin d'intégrer des patrons morphosyntaxiques à une combinaison telle celle qui sera envisagée, il est nécessaire d'affecter, à chacun d'entre eux, une grandeur numérique réelle (il ne faut pas se limiter à la décision binaire *« le patron capture ou ne capture pas »*

une expression). Des travaux tels que (Duclaye *et al.*, 2003) et (Vidrequin *et al.*, 2007) permettent cela en extrapolant une interprétation du rappel et de la précision. D'autres critères comme un critère de redondance (correspondant au nombre d'occurrences où un patron parvient à capturer une expression) pourrait aussi être pris en compte. Cependant, et selon nous, il s'agit d'un critère intervenant plutôt lors d'une étape de vérification qu'à celle d'extraction, d'autant que ce critère peut aussi être envisagé dans le cas des autres approches.

## 2.1 Formulation de la combinaison

En examinant le tableau 1, il ressort que l'expressivité de chacun des scores est différente. Ces différences peuvent (doivent) être perçues comme susceptibles de dégager une complémentarité des approches dont il est intéressant de tirer profit. Il s'agit justement de notre objectif principal : mettre en œuvre une combinaison de ces scores afin de capturer au mieux les critères envisagés (et peut être d'autres moins évidents). L'interrogation que nous avons porte sur la manière d'aboutir à une fusion optimale. Ainsi, comme cela a été déjà mentionné, ces scores proviennent de deux modules différents. Les premiers proviennent de celui en charge de la recherche de passage, les autres de celui en charge de l'extraction et de la sélection de la réponse. Ce constat nous amène à envisager une combinaison de ces scores en tenant compte de leur provenance, et à choisir une formulation de leur combinaison.

Ainsi, considérons *Cosine* et la compacité positionnelle. Elles apparaissent quasiment complémentaires (orthogonales), il est raisonnable de penser que leur combinaison peut être intéressante. Cette forme d'orthogonalité peut être rapprochée d'une notion d'indépendance, et faire penser à percevoir celles-ci comme l'expression de probabilités (cela même si elles n'en sont évidemment pas[4]). De plus, le calcul *Cosine* qui nous intéresse accepte une valeur comprise entre 0 et 1. Concernant la compacité, si nous avons déjà envisagé de l'exprimer d'une manière probabiliste (Gillard *et al.*, 2007), la compacité positionnelle n'est pas l'expression d'une probabilité. Cependant, nous l'amenons par une normalisation dans ce même intervalle de valeurs. Ainsi, il est possible d'envisager leur combinaison au travers d'un produit, justement dans le but de saisir cette indépendance / orthogonalité, soit :

$$scoreFinal = scorePassage \times scoreExtraction \quad (1)$$

Ou plutôt, afin de s'affranchir des problèmes de cadrage de valeur d'un produit de facteurs inférieurs à un :

$$scoreFinal = \log(scorePassage) + \log(scoreExtraction) \quad (2)$$

Du point de vue expérimental, lors d'EQueR, la combinaison retenue faisait également intervenir un décompte de Mots Communs (MC), dans le but de favoriser encore plus le cas où le nombre de ces MC dans les passages était optimal. Le score employé alors était :

$$scoreFinalAvecMC = \log(scorePassage) + \log(scoreExtraction) + \log(1 + MC) \quad (3)$$

Dans les expériences qui suivent, ces deux scores (2) et (3) sont envisagés, l'un étant une version avec décompte de mots communs (« avec MC »), l'autre sans (« sans MC »). Le score des passages sera tour à tour, un score de densité, de *Cosine*, de *Cosine+Okapi*, ou issu de *Lucene*. Dans le tableau 2, est également présentée une extraction sans aucune combinaison, soit depuis un décompte des mots communs, depuis un score de passage seul, ou depuis une compacité positionnelle, cela afin de vérifier le gain éventuel observé lors d'une combinaison.

---

[4] En effet, il n'existe pas de loi de probabilité avec une densité de probabilité, associée à ces scores, pas plus que de paramètres susceptibles de faire l'objet d'une procédure d'apprentissage comme par exemple via l'estimateur du maximum de vraisemblance.

## 2.2 Expérimentation de la combinaison

Cette partie présente les résultats obtenus en fin de chaîne par le système de QR aux 400 questions factuelles d'EQueR. Notre objectif est d'examiner l'impact des deux combinaisons envisagées lors de nos participations à EQueR et à QA@CLEF-2006. La première repose sur un score de passage, un score de compacité et un décompte des mots communs (*« avec MC »*, $b_1$/, comme lors de notre participation à EQueR) ; la seconde est sans décompte des mots communs (*« sans MC »*, $b_2$/, et correspond à notre participation à CLEF-2006). De plus, lors de ces campagnes nous avions fait le choix a priori de nous limiter à des passages de *trois* phrases. Ces résultats nous permettent d'examiner ce choix *a posteriori*.

Les évaluations sont obtenues au moyen d'une procédure automatique reposant sur une référence que nous avons constituée à partir des réponses des participants de la campagne. Deux déclinaisons, stricte et tolérante sont envisagées et constituent des bornes inférieures et supérieures. Par ailleurs, il est à noter que, dans une autre expérience, nous avons pu comparer les sorties obtenues depuis cette procédure automatique avec celles faites par un évaluateur humain et nous avons constaté que les différences entre les jugements humains/automatiques apparaissaient acceptables et comprises entre *-3,5%* à *+1%* suivant le type d'évaluation.

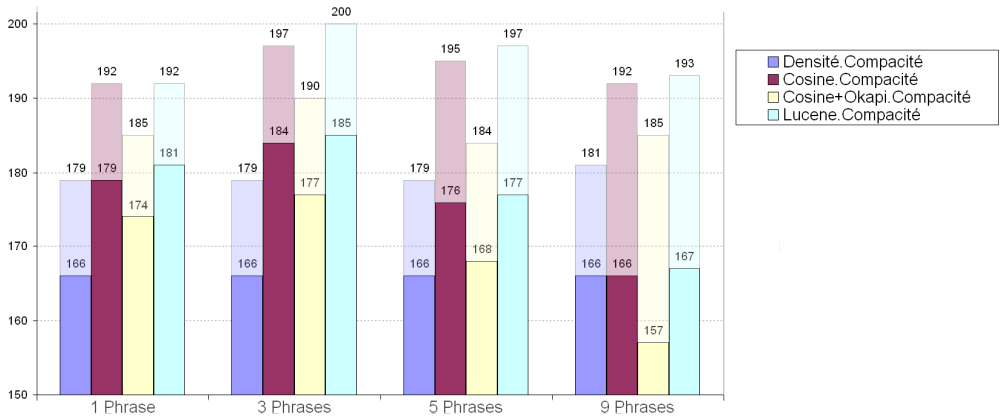
Les nombres présentés dans le tableau 2 (voir en fin d'article) sont le nombre de réponses correctes précises (aussi dites *« exactes »*, en ce sens que la *« chaîne de caractères »* correspondante contient uniquement les informations utiles pour répondre, sans oubli, ni ajout) trouvées suivant le type d'évaluation (stricte ou tolérante) en considérant la réponse proposée par le système au premier rang pour chacune des questions factuelles de la campagne EQueR. Ce type d'évaluation sur une unique réponse autorisée par question correspond à la définition de l'évaluation CLEF-2006 (il est à noter que, lors d'EQueR, les cinq premiers rangs étaient examinés, aussi le graphique 1, plus loin, présente-t-il des résultats lorsque la réponse peut être trouvée parmi les cinq premières pour un cas qui nous intéresse particulièrement : celui des combinaisons entre les scores de passages et notre score d'extraction de *compacité*).

**Bilan du tableau 2** (voir en fin d'article)**.** D'une manière globale, on peut noter que les performances des combinaisons densité et compacité sont assez indépendantes de la dimension des passages. Cela tend à prouver que la complémentarité des deux approches est moindre. Cela s'explique par le fait que les principes sous-jacents à leur définition sont identiques et reposent sur des notions de proximité (et peut-être de symétrie, puisque la phrase intéressante du point de vue de la densité occupe la position centrale dans le passage). La compacité apparaît comme un choix intéressant même lorsqu'elle est employée seule. Elle se situe d'ailleurs juste derrière les combinaisons scores de passages et compacité du point de vue des performances. Pour la compacité employée hors d'une combinaison, les meilleurs résultats sont obtenus depuis des passages de *neuf* phrases (*[104..118]*), lorsque les passages sont préalablement filtrés par notre score de densité, et de *trois* phrases sinon (*[102..118]*).

La meilleure combinaison est obtenue par l'emploi d'une combinaison entre *Lucene* et compacité pour des dimensions de passages d'*une* ou de *cinq* phrases (respectivement *[121..135]*, *[120..136]*), ou compacité et *Cosine* sur *trois* phrases (*[120..136]*). Concernant la combinaison avec mots communs, notre souhait initial consistant à favoriser les combinaisons avec un grand nombre de mots communs s'avère être une mauvaise chose, au moins en l'état : peut-être manque-t-il des coefficients adaptés dans l'expression de la combinaison linéaire ?

Enfin, nous avons constaté qu'il existe une variabilité non négligeable entre les différentes méthodes sur les questions répondues (du point de vue des pertes et gains) et cela amène à supposer que ces expérimentations peuvent donner lieu à d'autres améliorations en fin de traitement : il doit être possible de maximiser encore les gains tout en minimisant les pertes.

Notre principal objectif sur la combinaison de scores apparaît rempli : une combinaison entre un score de compacité et un score *Lucene*, ou *Cosine*, permet un gain d'environ *+18%* sur les meilleures autres configurations sans combinaison (qui proviennent par ailleurs de la compacité que nous avons définie), cela si l'on observe la réponse proposée au premier rang.




*Graphique 1. Nombre de réponses précises correctes trouvées parmi les cinq premières depuis des combinaisons entre différents scores de passages et score de compacité sur le jeu des 400 questions factuelles de la campagne EQueR (premier plan évaluation stricte, second plan évaluation tolérante. Note : ce graphique correspond à la ligne « sans MC » ($b_2$) du tableau 2 avec une réponse correcte trouvée parmi les cinq premières plutôt qu'uniquement au premier rang).*

Une autre interrogation qui apparaissait en filigrane concernait l'influence de la dimension des passages sur les performances de notre système. Il ne ressort pas une préférence nette lorsqu'une réponse au premier rang est considérée (soit les résultats présentés dans le tableau 1). En revanche, l'examen du nombre de réponses correctes trouvées parmi les cinq premières (voir graphique 1) nous permet de mettre en évidence que notre choix *a priori* de se limiter à un contexte de *trois* phrases correspond à une configuration acceptable. En outre, cela permet également de constater que la combinaison *Cosine* et compacité ou *Lucene* et compacité produit dans les cinq premiers rangs, là encore, les meilleurs résultats.

# 3 Conclusion et perspectives

Dans cet article, nous avons étudié une combinaison des scores intervenant dans la sélection de passages et l'extraction d'une réponse. Notre objectif était d'utiliser cette combinaison pour profiter pleinement de la complémentarité des scores employés. Nous avons notamment examiné la complémentarité entre le score de compacité positionnelle, que nous avons défini pour la sélection et l'extraction d'une réponse, avec des méthodes de recherche vectorielle conventionnelle ou notre score de densité.

Comme attendu, nous avons pu constater un gain lorsqu'un score de passages et de compacité étaient combinés. Nous avons mis en évidence que la meilleure complémentarité s'obtenait depuis un score de passage provenant d'une similarité *Cosine* ou *Lucene*. Les améliorations de performances, par rapport à notre score de compacité employé seul, sont d'environ *+18%* lorsque seulement la première réponse est considérée, et entre *+10* et *+15%* lorsque les cinq premières le sont. Ces hausses, d'au moins *+10%*, nous permettent de conclure, malgré les précautions à prendre vis-à-vis d'évaluations automatiques en QR, que ces améliorations doivent être effectives (et non liées à des artefacts d'évaluation lorsque cette dernière est réalisée par des patrons et identifiants de documents). Enfin, puisque les résultats présentés ici sont ceux en fin de chaîne de traitement, nous avons pu mettre en évidence les bonnes performances de la compacité employée seule, mais également que sa combinaison avec le

score de densité était relativement stable, ce qui s'explique par la grande similitude entre les deux approches. Parallèlement, nous avons vérifié *a posteriori* qu'un découpage en passages de *trois* phrases correspondait à un optimum local pour notre système de Questions/Réponses, alors qu'il était un compromis technique *a priori*.

Une étude supplémentaire intéressante à mener serait d'étudier la distribution et les valeurs numériques engendrées par la combinaison proposée afin de mieux cerner les éventuels effets de bord liés à ces valeurs numériques (cas d'égalité, de *« décrochage »*, *etc.*). Les autres perspectives sont nombreuses : il reste à dégager des critères dépendants des grandes classes de questions ou même, au niveau des questions elles-mêmes (distribution des mots dans les passages, mise en évidence des mots avec les meilleures contributions numériques, *etc.*). Ces critères, ou d'autres nouveaux scores, pourraient alors être intégrés dans une combinaison telle que proposée. Par ailleurs, nous avons pu constater, lors de comparaisons deux-à-deux, une rotation d'une quinzaine de réponses qui sont trouvées ou sont perdues suivant les expériences. Les critères pour aller chercher ces réponses sont à déterminer. Enfin, comme la combinaison que nous avons testée est uniquement numérique, il pourrait être opportun, avec un corpus adéquat, d'effectuer un apprentissage de coefficients pour maximiser les performances de cette combinaison et tenir compte des dynamiques peu visibles.

**Remerciements.** Les auteurs tiennent à remercier le Dr. Olivier Ferret pour sa lecture attentive de cet article ainsi que ses très utiles commentaires et suggestions.

# Références

AYACHE C., CHOUKRI K., GRAU B., (2005) Rapport de la Campagne EVALDA/EQueR Evaluation en Questions-Réponses http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf

DUCLAYE F., COLLIN O., YVON F., (2003). Apprentissage Automatique de Paraphrases pour l'Amélioration d'un Système de Questions-Réponses. Actes de *TALN* 2003, 115-124.

FERRET F., GRAU B., HURAULT-PLANTET M., ILLOUZ G., MONCEAUX L., ROBBA I., VILNAT A., (2002). Recherche de la réponse fondée sur la reconnaissance du focus de la question. Actes de *TALN*, Nancy, 307-316.

GILLARD L., BELLOT P., EL-BÈZE M., (2007). D'une compacité positionnelle à une compacité probabiliste pour un système de Questions/Réponses. Actes de la *4ième COnférence en Recherche d'Informations et Applications (CORIA) 2007*, 28-30 mars 2007, Saint-Etienne (France), 271-286.

GILLARD L., SITBON L., BELLOT P., EL-BÈZE M., (2005). Dernières évolutions de SQuALIA, le système de Questions/Réponses du LIA. Dans Réponses à des questions, Traitement Automatique des Langues (TAL), 2005, Hermès, Paris (France), volume 46, n°3/2005, 41-70.

LUHN H.P., (1958). The Automatic Creation of Literature Astracts. *IBM Journal of Research and Development*, Volume 2, Issue 2, avril 1958, 159 165.

VIDREQUIN C., EL-BÈZE M., TORRES-MORENO J.M., SCHNEIDER J.J., (2007). Génération et enrichissement automatique de listes de patrons de phrases pour les moteurs de questions-réponses. Actes des *Cinquièmes journées Extraction et Gestion des Connaissances (EGC'2007)*. Namur (Belgique). 23-26 janvier 2007. 207-208.

TELLEX S., KATZ B., LIN J., MARTON G., FERNANDES A., (2003). 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, 41-47.

VOORHEES E.M., HARMAN D., (2005) *TREC Experiment and Evaluation* Dans *Information Retrieval*. MIT Press 2005, chapitre 10. 233-257.

WILKINSON W., J. ZOBEL J., SACKS-DAVIS R., (1996). Similarity measures for short queries. Actes de *The 4th Text Retrieval Conference (TREC-4)*. Gaithersburg, Maryland (USA). NIST special publication 500-236 (novembre 1995), 1996. 277-286.

| | ***Scores de passages mis en œuvre*** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ***Densité*** | | | | ***Cosine*** | | | | ***Cosine+Okapi*** | | | | ***Lucene*** | | | |
| *Évaluation* | *stricte* | | *tolérante* | | *stricte* | | *tolérante* | | *stricte* | | *tolérante* | | *stricte* | | *tolérante* | |
| ***a/*** | ***Comparaisons entre décompte des mots communs et score de passage*** | | | | | | | | | | | | | | | |
| *Mots communs* | 48 (1p) | 43 (3p) | 57 | 50 | 47 | 37 | 57 | 45 | 47 | 36 | 57 | 44 | 47 | 39 | 57 | 47 |
| *(MC)* | 41 (5p) | 25 (9p) | 47 | 32 | 29 | 15 | 41 | 31 | 29 | 16 | 41 | 32 | 30 | 15 | 42 | 30 |
| *Scores de passages* | 66 | 50 | 74 | 54 | 64 | 48 | 77 | 55 | 71 | 48 | ***84*** | ***61*** | 75 | 48 | 87 | 64 |
| *(à lire en colonne)* | 50 | 29 | 54 | 31 | 38 | 26 | 49 | 34 | 39 | 23 | ***49*** | ***34*** | 36 | 23 | 51 | 33 |
| *Compacité* | 100 | 100 | 117 | 117 | 101 | 102 | 118 | 117 | 101 | 102 | 118 | 102 | **101** | **102** | 118 | 117 |
| | 102 | 104 | 118 | 118 | 98 | 89 | 119 | 116 | 98 | 89 | 119 | 120 | **98** | **89** | 119 | 119 |
| ***b/*** | ***Comparaisons entre deux combinaisons impliquant des scores de passages et de compacité, avec ou sans décompte des mots communs (MC)*** | | | | | | | | | | | | | | | |
| $b_1$/ *avec MC* | 96 | 86 | 104 | 95 | 98 | 81 | 109 | 95 | 97 | 80 | 108 | 93 | 98 | 85 | 109 | 101 |
| *(formule 3 du 2.1)* | 85 | 82 | 94 | 91 | 84 | 65 | 100 | 95 | 79 | 64 | 95 | 90 | 86 | 69 | 102 | 98 |
| $b_2$/ *sans MC* | 104 | 103 | 119 | 118 | **113** | **120** | 128 | 136 | 115 | 115 | 127 | 131 | 121 | 118 | 135 | 134 |
| *(formule 2 du 2.1)* | 105 | 106 | 120 | 119 | **114** | **106** | 134 | 141 | 111 | 103 | 128 | 128 | 120 | 113 | 136 | 140 |

*Tableau 2. Diverses comparaisons entre des méthodes « séquentielles » (a : décompte des mots communs, scores de passages et score de compacité) et deux combinaisons de ces méthodes (b) en vue d'une extraction et de la sélection d'une réponse finale. Les passages dans lesquels a lieu l'extraction sont constitués d'une, trois, cinq ou neuf phrases (à lire suivant le sens de lecture de gauche à droite dans une même case). Les nombres présentés correspondent au nombre de réponses précises correctes à une question factuelle au premier rang suivant la nature des évaluations : strictes ou tolérantes. Les questions considérées sont les 400 questions factuelles de la campagne EQueR.*

*Exemples de lecture : premier exemple : pour une sélection des passages depuis Lucene (dernière colonne), suivi d'une sélection des réponses par un score de compacité (troisième ligne) et dans le cas d'une évaluation stricte, soit les valeurs en gras et soulignées dans le tableau, le nombre de réponses précises correctes trouvées au premier rang est de 101 si les passages de départ étaient constitués d'une seule phrase, 102 pour trois phrases, 98 pour cinq phrases, et 89 pour des passages de neuf phrases.*

*Deuxième exemple : pour une sélection des passages depuis un score de passage Cosine+Okapi (pour les scores de Passages présentés en seconde ligne, la nature du score doit être lu depuis le titre de la colonne seulement) suivi d'une sélection depuis une stratégie sur la première occurrence du type de réponse compatible avec celui recherché, et dans le cas d'une évaluation tolérante, ce qui correspond aux (valeurs en gras et italique dans le tableau, le nombre de réponses correctes trouvées au premier rang est de 84 si les passages de départ étaient constitués d'une seule phrase, 61 pour ceux de trois phrases, 49 pour ceux de cinq phrases, et 34 pour des passages de neuf phrases.*

*Troisième exemple (valeurs en gras simple dans le tableau) : pour une sélection des passages depuis un score de passage Cosine (deuxième colonne), suivi d'une sélection des réponses faisant intervenir la combinaison entre un score de compacité et de passage (ligne $b_2$/ « sans MC ») et dans le cas d'une évaluation stricte, le nombre de réponses correctes trouvées au premier rang est de 113 si les passages de départ étaient constitués d'une seule phrase, 120 pour ceux de trois phrases, 114 pour ceux de cinq phrases, et 106 pour les passages de neuf phrases.*