D'une compacité positionnelle à une compacité probabiliste pour 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 présentons une discussion sur la définition d'un score de compacité pour permettre l'extraction d'une réponse dans un système de Questions/Réponses. Ce score de compacité qui peut être succinctement décrit comme une fonction liée a la densité des termes de la question dans le voisinage d'une réponse candidate, est présenté en détail. Ensuite, une discussion nous amène à envisager une extension de ce score, initialement défini d'un point de vue positionnel, vers un modèle probabiliste ; cela afin de mieux prendre en compte des critères d'importance variable pour les mots de la question. Cette discussion est validée par des expériences où ce critère d'importance est modélisé à l'aide d'un calcul d'IDF (« Inverse Document Frequency »), et nous montrons que nous obtenons des résultats similaires à ceux obtenus lors de notre participation à la campagne d'évaluation EQueR des systèmes de Questions/Réponses.*

ABSTRACT. *In this paper, we present an answer extraction algorithm based on a density ("compactness") score for a question-answering system. This compactness score is first described from a positional point of view. Next, a discussion will bring us to envision an extension of this score towards a probabilistic model in order to better take into consideration question words importance by using theirs Inverse Document Frequencies (IDF). We experiment these positional and probabilistic compactness on a subset (composed by factual questions) of the French question answering campaign EQueR, and show that this IDF extension is worth interest.*

MOTS-CLÉS : *Système de questions réponses, extraction d'une réponse, compacité, IDF, densité, EQueR.*

KEYWORDS: *French Question answering, answer extraction, inverse document frequency, compactness density.*

CORIA 2007, pp. 271–286
Saint-Étienne, 28–30 mars 2007

1. Introduction

Les systèmes de Questions/Réponses (QR) se proposent d'extraire, depuis une masse de documents, *LA* réponse à une question formulée en langage naturel. En cela, ils s'opposent, ou plutôt complètent, les systèmes de recherche d'information classiques puisque ces derniers laissent le soin de cette recherche d'information *précise* à leur utilisateur après en avoir limité la recherche à quelques documents susceptibles de contenir cette réponse. Ainsi, comme il est possible de le percevoir aisément, la tâche de répondre à une question n'est pas triviale, surtout si celle-ci peut être ouverte et porter sur l'intégralité des connaissances humaines. Aussi, les systèmes de Questions/Réponses (sQR) actuellement à l'étude s'intéressent à des types de questions relativement particuliers, et principalement à des questions dites factuelles : c'est-à-dire dont la réponse peut être un court énoncé, et plus spécifiquement l'expression d'une information sémantique précise telle que le nom d'une personne, la date d'un événement, une valeur numérique, un lieu, ou même la raison pour laquelle une personne est connue. De plus, cette réponse est obtenue par extraction, plutôt que par synthèse (la réponse est une bribe de texte), depuis un ensemble de documents typiquement de type journalistique. Enfin, l'étude des sQR est encouragée au travers de différentes campagnes d'évaluation à grande échelle : la première de ces campagnes qui a été conduite dans le cadre du cycle des *« Text REtrieval Conference »* (TREC ; Voorhees & Harman, 2005), portait sur la langue anglaise et, elle est, depuis, reconduite et complexifiée d'année en année. La campagne QA@CLEF (pour *« Question Answering at the Cross Evaluation Language Forum »*) s'intéresse aux langues européennes et plus particulièrement à des systèmes multilingues où la réponse est recherchée dans une langue différente de celle dans laquelle la question est formulée. Les langues asiatiques sont étudiées dans le cadre des pistes QR des ateliers NTCIR. Et, en 2004, dans le cadre de l'action Technolangue, la campagne Évaluation en Questions-Réponses (EQueR) a été menée sur des QR en français.

Ce contexte préalable étant posé, le présent article s'intéresse à l'étape finale d'un système de Questions/Réponses : à savoir, l'extraction d'une réponse lorsque le type de l'information recherchée pour une question est identifié et qu'il existe différents candidats de réponses envisageables et localisés, au travers d'un balisage particulier, à l'intérieur d'un ensemble de passages.

Aussi, ce qui nous intéresse tout particulièrement est la possibilité de modéliser un calcul numérique permettant de choisir au mieux une réponse parmi d'autres (et par conséquent qui apparaissent dans des contextes différents) sachant les mots rencontrés dans une question. Pour ce faire, un score que nous appelons *« Compacité »* est défini. Cette compacité s'apparente à un calcul de densité normalisée sur la présence des mots de la question au voisinage d'une réponse candidate. Elle a été intuitivement définie selon un paradigme positionnel, et a été utilisée lors de nos deux participations à des campagnes d'évaluation des systèmes de QR. Après une présentation de cette *« compacité positionnelle »*, nous

envisageons une extension vers une *« compacité probabiliste »* ; cela afin de prendre en compte une notion d'importance des mots utilisés pour préférer une réponse à une autre. Dans cet article, ce critère d'importance est modélisé au travers d'un calcul d'IDF (*« Inverse Document Frequency »*). Les résultats expérimentaux montrent que cette nouvelle définition obtient des performances comparables à notre précédente définition positionnelle tout en ouvrant des perspectives et notamment une meilleure prise en compte d'éventuels liens sémantiques comme l'hyperonymie en vue de leur intégration dans le calcul d'un score numérique.

Il est à noter que ce travail reprend une partie de (Gillard *et al.*, 2006a) puisque la compacité positionnelle y était présentée, cependant il complète cette définition d'alors par des exemples de calcul mais c'est surtout l'extension à un modèle probabiliste qui en fait son originalité.

La suite de cet article est découpée comme suit : afin d'éclairer les points qui pourraient rester obscurs sur le fonctionnement d'un système de Questions/Réponses, la section 2 présente, sur un exemple, et au travers d'un schéma, le processus de réponse pour un sQR générique. La section 3 présente brièvement quelques autres approches connexes pour l'extraction finale d'une réponse. La section 4, concerne l'essentiel de la discussion sur la compacité, sa définition et son extension probabiliste. Enfin, la section 5 détaille le protocole expérimental basé sur notre participation à la campagne EQueR (Gillard *et al.*, 2004), ainsi que les résultats qui montrent que nos compacités obtiennent des performances similaires.

2. Architecture générique pour un système de Questions/Réponses

Avec l'intérêt grandissant pour la problématique Questions/Réponses, de nombreux systèmes de QR expérimentaux ont été développés. Au delà de leurs différences intrinsèques, il est possible de dégager une architecture générique et des processus communs. Ainsi, un système de Questions/Réponses peut être schématiquement décrit au travers d'un enchaînement de trois étapes principales : une analyse de la question, un traitement des documents et 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 (avec un Système de Recherche Documentaire, SRD) suivie d'une exploitation des documents sélectionnés à l'issue de cette recherche. Enfin, d'un point de vue logiciel, chacune des étapes considérées donne lieu à un composant éponyme.

La figure 1 présente cette architecture schématique à trois, et plus généralement quatre composants mais également leurs principales interactions. En effet, chaque étape produit en sortie des informations utilisées en entrée de l'étape suivante (voire en entrée de plusieurs de celles situées en aval). Cependant, pour quelques systèmes, plus complexes, il manque également à cette figure les éventuels flux

supplémentaires générés lors d'allers-retours et/ou rétroactions entre les différents composants (relaxation de contraintes, etc.).

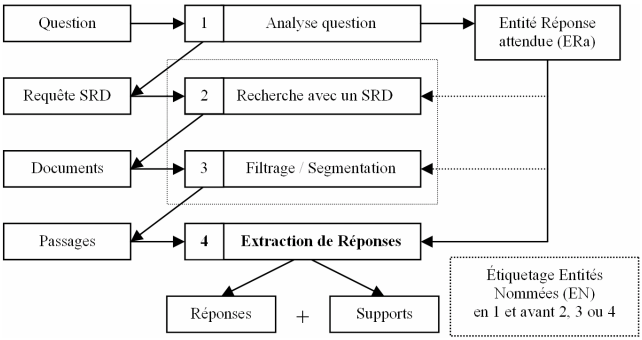


Figure 1. *Architecture schématique d'un système de QR*

À titre d'illustration, le déroulement de cette architecture générique pour la question formulée en langage naturel *« En quelle année est né Nelson Mandela ? »* (provenant de la campagne *CLEF-2004/0065*) pourrait être le suivant :

- L'étape d'analyse d'une question associe une étiquette de la nature de l'information qui va être recherchée (appelée également Entité, ou Type, Réponse attendue), soit ici, l'*année* associée à une *date*, et plus spécifiquement l'*année* d'une *date de naissance* ; elle construit également une requête à destination d'un système de recherche documentaire, par exemple, depuis les lemmes obtenus grâce à une analyse morphosyntaxique et après un filtrage des pronoms interrogatifs et des mots vides de sens (*« en, quel, année, être »*), la requête est alors constituée de *« naître »*, *« Nelson »*, et *« Mandela » ;*

- le système de recherche documentaire propose une liste des identifiants des documents du corpus qui contiennent tous ces lemmes, soit les documents *« ATS.940426.0081 », « LEMONDE94-000873-19940509 », « LEMONDE94-003506-19940429 », « LEMONDE95-000054, LEMONDE95-012364 »*. Bien que cette liste ne soit pas ordonnée, elle pourrait l'être selon un indice de similarité prenant en compte une fréquence d'apparition des lemmes de l'étape précédente ;

- le composant en charge du traitement des documents effectue un étiquetage morphosyntaxique des documents présélectionnés. Il procède également au balisage des entités de type *date* à l'intérieur de ceux-ci. Il découpe ces documents en bloc ou passages et retire 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, une *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 blocs précédemment conservés. Et finalement, dans ce dernier cas, le *« 1918 »* de *« Nelson Mandela est né le [DATE18 juillet 1918] dans un village xhosa du Transkeï. » (ATS.940426.0081)* apparaîtrait comme une meilleure réponse que le *« 1912 »* rencontré dans le passage suivant : *« Nelson Mandela _ soixante-quinze ans _ n'était pas né quand le Congrès national africain (ANC) vit le jour ([DATE1912]). » (LEMONDE94-003506-19940429)*.

C'est sur cette dernière étape (en gras sur la figure 1) que porte la suite de cet article et plus spécifiquement sur des calculs permettant de préférer l'une des réponses aux autres et par conséquent de l'extraire.

4. D'une compacité positionnelle à une compacité probabiliste

Cette partie propose et discute les différentes réflexions qui nous ont conduits à la définition des deux compacités qui sont expérimentées en section suivante, et tout particulièrement celles relevant de la compacité probabiliste. En effet, cette dernière nous paraît un modèle original et adapté à notre tâche, tout en autorisant une extension afin de prendre en compte des relations sémantiques[1] entre les mots.

Le problème que nous posons peut être exprimé de la façon suivante : comment est-il possible de choisir la réponse à une question, sachant que nous disposons des informations suivantes :

- Cette question contient évidemment des mots qui sont autant d'indices permettant d'y répondre. De plus, elle est associée à une étiquette sémantique de la nature de l'information qui doit être recherchée (nous l'appelons une Entité Réponse attendue ou *ERa*) qui permet de contraindre les candidats potentiels ;
- La recherche à effectuer a lieu dans un ensemble de passages contenant au moins une Entité Réponse candidate (*ERc*) d'un type sémantique compatible avec celui associé à la question (l'*ERa*). Cette Entité Réponse candidate est composée d'un mot ou d'un groupe de mots. Elle est identifiée, par exemple, grâce à un balisage tel que ceux obtenus après un étiquetage en Entités Nommées. Ces passages contiennent aussi, et très vraisemblablement, des mots en commun avec ceux de la question.

Ainsi, à partir de ces hypothèses, nous cherchons à définir un score susceptible de départager les différentes Entités Réponses candidates.

[1] Cependant, à ce point, nous supposons ne pas disposer de ces connaissances de plus haut niveau qui permettent une compréhension « plus sémantique » de la question et des passages.

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 repère[2] gradué par la position des mots, 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 incorrects. En effet, ils sont sources de bruit et nuisent à la qualité de la relation au sein du couple < *Question et Réponse candidate* >.

Cependant, un tel 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 entre-elles les réponses à une même question[3]. 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.

Au final, 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 Réponse candidate. Cela rejoint également la proposition de (Radev *et al.*, 2002) : *« Phrasal answers tend to appear near words from the query »*.

Formalisons : soit *QW* l'ensemble des mots (éventuellement non vides) de la question, soit *w* l'un de ces mots, soit ERc_i une Entité Réponse candidate *i* ; la Compacité associée à une Réponse candidate se définit comme (les |barres| représentent le cardinal d'un ensemble) :

$$compacité(ERc_i) = \frac{\sum_{w \in QW} contribCompacité(w)_{ERc_i}}{|QW|} \quad [1]$$

Ce qui s'exprime encore comme *« la somme des contributions à la compacité des mots de la question entourant cette Réponse candidate divisée par le nombre de mots de cette question »*.

La contribution à la compacité d'un mot *w* correspond à la précision à l'intérieur d'une fenêtre centrée sur l'Entité Réponse candidate ERc_i pour les mots de la

[2] D'où l'utilisation de l'adjectif « positionnelle » dans l'appellation Compacité positionnelle.
[3] Et non pas l'ensemble des réponses à toutes les questions, ce qui était demandé lors de TREC-11.

question à l'intérieur d'un rayon R fixé par l'occurrence la plus proche X_p du mot X (cf. algorithme 1 et [2]).

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

$$contribCompacité(w)_{ERc_i} = \frac{|Z|}{2R+1} \quad [2]$$

Algorithme 1. *Calcul de la contribution à la compacité d'un mot w.*

Il est à noter que dans ces formulations, le rayon R pourrait plutôt être fixé par l'occurrence X_{max} qui maximiserait le calcul de la contribution à la compacité. Une autre variation serait de ne pas faire le choix d'une fenêtre centrée, mais plutôt de la positionner seulement à gauche ou seulement à droite de l'ERc_i (qui jouerait alors le rôle d'une extrémité de cette fenêtre) ; dans ce cas, il faudrait ajuster le facteur de normalisation présent au dénominateur de [2] à la dimension de cette fenêtre.

Afin d'éclairer sur les formulations ci-dessus, nous poursuivons par un exemple simplifié d'un calcul de compacité, avec les variations évoquées : le regard se porte seulement à gauche OU seulement à droite d'une *ERc*, aussi le facteur de normalisation de la contribution à la compacité est-il de $R+1$ (Contre $2R+1$ dans la formule 2 qui correspond à une fenêtre centrée et un regard des deux cotés).

Soit la configuration suivante : après nettoyage la question traitée ne comporte plus que 4 mots $\{w_1, w_2, w_3, w_4\}$, le passage contient une *ERc* et certains de ces mots :

w_3		w_1	ERc_i		w_3			w_4
-3	-2	-1	0	1	2	3	4	5

Calcul de la contribution de w_1: le mot w_1 apparaît une seule fois, à une distance de *1* de l'*ERc* : l'ensemble *Z* est $\{w_1, ERc\}$, aussi sa contribution est de $(1+1)$ sur une distance de *2* mots $(R+1)$, ce qui donne une contribution de $(1+1)/2=1$.

Le mot w_2 n'apparaît pas dans la phrase ; par conséquent il a une contribution nulle à la compacité.

Le mot w_3 apparaît deux fois dans la phrase, la première fois avant, à une distance de *3* et la seconde fois après, à une distance de *2*, aussi, y a-t-il deux contributions envisageables, comme cela a été dit, il est nécessaire de définir une stratégie telles que le choix de prendre en compte la plus proche occurrence, ou celui de considérer la meilleure contribution. Considérons par exemple, ce dernier cas. La

première apparition de w_3 (position *-3*) correspond à *{w_3, w_1, ERc}* soit *3* mots utiles sur une distance de *4*, la contribution est de *3/4*. La seconde apparition (position *2*) correspond à *{ERc, w_3}* sur une distance de *3*, soit *2/3*. Ainsi, c'est la première apparition qui apporte la meilleure contribution bien qu'elle soit la plus éloignée de l'*ERc*.

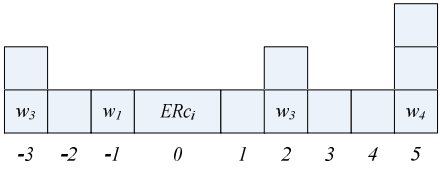
Enfin, le mot w_4 apparaît une seule fois en position *5*, la séquence est *{ERc, w_3, w_4}*, sur une distance de *6*, soit *3/6* ou *1/2*.

La somme des contributions est *(1+0+3/4+1/2)* divisée par *4*, le nombre de mots de la question. Le score final de la compacité pour cet exemple est de *9/16≈0.5625*.

Les calculs de compacité ainsi présentés ont été mis en œuvre lors de nos expériences et participations aux campagnes EQueR et CLEF 2006 (Gillard *et al.*, 2006b) à la différence que la plus proche occurrence a été retenue pour le calcul du rayon de la fenêtre (comme explicité dans l'algorithme 1) ; et cette fenêtre a été centrée avec un regard à gauche et à droite (aussi la normalisation s'effectue par *2R+1* comme en formule 2). Les performances obtenues étaient acceptables par rapport aux autres systèmes participants : elles sont, sur les seules questions factuelles, d'environ 29% de bonnes réponses pour EQueR et 37% lors de CLEF 2006, ce qui positionne notre système dans les 3 premiers. Cela nous permet de supposer que ces calculs correspondent à une extraction proche de l'état de l'art.

Pourtant dans ces calculs, le fait de donner à tous les mots de la question une importance équivalente, peut apparaître comme un défaut. En effet, tel ou tel mot devrait pouvoir peser plus qu'un autre sur le calcul et améliorer la contribution globale de l'ensemble. Cette notion d'importance peut naturellement être rapprochée de celle de la « quantité d'information » qu'il est possible d'affecter à un mot, ou, dans le modèle vectoriel classique, de la notion d'IDF (« *inverse document frequency* »). Cela aurait pu aussi être envisagé d'un point de vue syntaxique : dans un couple < *nom, suivi d'un adjectif* >, la présence du nom peut apparaître plus importante que celle de l'adjectif qui le complète.

Reprenons le même exemple avec différentes importances pour les mots de la question : w_1 et w_2 ont une importance évaluée à *1*, w_3 est plus important avec une valeur de *2*, enfin w_4 a une importance de *3*. La somme des importances est de *7*. Dans ces calculs, l'importance du mot en cours de considération est prise en compte à sa valeur propre alors que les autres le sont avec une valeur de *1*.



Le calcul donne : (là encore, les cases vides de mots importants sont considérées avec une importance de 0 puisque des mots vides de sens ou certaines ponctuations n'apportent pas « d'éloignement sémantique » bien qu'ils contribuent à un « éloignement positionnel ». Cependant, il ne faudrait pas perdre de vue le problème de la prise en en compte d'une ponctuation aussi forte que le point final *« . »* surtout dans le cas ou aucune anaphore n'existe entre deux phrases consécutives. Par ailleurs, afin d'être cohérent avec notre proposition de départ, le facteur de normalisation, dans ce cas n'est plus le nombre de mots de la question mais bien la somme des importances maximales envisageables).

Contribution de w_1 : *(1+1)* sur une distance de *2*, ce qui amène à une contribution de *2/2=1*,

w_2 : contribution nulle puisque le mot n'apparaît pas

w_3 : pour la première apparition : *{w_3, w_1, ERc}* avec une importance de *(2+1+1)=4*, puisque celle de w_3 est de *2*, et qu'un seul autre mot (w_1) est présent ; à diviser par une distance de *3*, soit *4/3≈1.33* (à rapprocher de *3/4* dans le cas précédent). Et pour la seconde apparition *{ERc, w_3}* avec une importance de *(1+2)/3=1* (*2/3* dans le cas précédent).

w_4 : Pour atteindre w_4, *{ERc, w_3, w_4}* sont rencontrés. L'importance de w_3 n'entre en compte qu'à hauteur de *1* puisque l'intérêt porte sur w_4, qui a lui une importance de *3*, aussi la contribution de w_4 est de *(1+1+3)=5* à diviser par *5*, soit *5/5=1*.

Le score final de compacité s'établit à *(1+0+4/3+1)=10/3* à normaliser par la somme des importances (*7*), soit à peu près *0.47619*.

Nous venons de voir qu'il est possible de choisir une Entité Réponse dans un contexte favorable au sens de *« qui partage dans son voisinage un certain nombre de mots communs avec ceux présents dans la question »*.

Cependant, il est évident que cette façon de procéder ne permet pas de prendre la mesure de toutes les subtilités de la langue naturelle : ne sont considérés que les mots communs à la fois à la question et à la phrase. Par contre, dans le cas particulier où aucun des mots de la question n'apparaîtrait dans la phrase mais des synonymes de chacun d'entre eux *{w_1', w_2', w_3', w_4'}*, les calculs de compacité ainsi définis ne permettent que d'obtenir un score nul. Aussi, il serait peut-être intéressant de modéliser la notion de synonymie au travers d'un coefficient comme celui d'importance. Il en va de même pour les hyperonymes qui apparaissent comme des remplaçants possibles de leur hyponyme, mais probablement avec une importance moindre. Enfin, dans le cas très particulier des antonymes et des négations, ceux-ci pourraient apporter un *« éloignement supplémentaire »* dans les distances et par la même pénaliser les éventuelles contributions qui les engloberaient.

Ces différents constats étant faits, l'intérêt pour la compacité apparaît multiple : capturer au travers d'une proximité des mots de la question dans le voisinage d'une Entité Réponse candidate, la meilleure d'entre elles. Mais il faudrait prendre en

compte, d'une certaine façon et au mieux, les éventuels problèmes classiques des différences de vocabulaire (synonymes, *etc.*) et de leur importance relative.

Cette approche peut nous amener à raffiner le modèle déjà défini d'un point de vue positionnel (notion d'axe dont l'origine est l'*ERc*) vers un modèle probabiliste.

En revenant au cas le plus simple : ce qui apparaît intéressant est la présence des mots de la question dans une phrase ou d'une manière plus générique un segment. Cependant, il apparaît intéressant d'envisager un degré d'importance différent pour chacun d'entre eux. (Sparck Jones *et al.*, 1998) proposent, pour une recherche documentaire classique, de considérer une contribution qui dépend du nombre de document dans lequel un terme (au sens le plus générique d'objet tel qu'un mot ou une expression) apparaît et du nombre de documents dans la collection. Par analogie, pour le cas de segments, la *« Collection Frequency Weight »* (désormais plus connu sous l'acronyme d'IDF pour *« Inverse Document Frequency »*) serait :

$$CFW(w_i) = IDF(w_i) = -\log \frac{n_i}{N_S} \qquad [3]$$

avec n_i le nombre de segments contenant le mot w_i et N_s le nombre de segments total.

Ainsi, si P_i est la probabilité associée à la variable aléatoire correspondant au fait d'avoir rencontré le mot w_i au moins une fois dans un segment. Depuis une telle présence (KSJ parle d'*incidence*), la probabilité sous-jacente peut être estimée par :

$$P(w_i) = \frac{n_i}{N_S} \qquad [4]$$

Ce qui correspond à une probabilité qui favorise les termes les plus fréquents dans un segment donné. Une autre façon d'exprimer la rareté d'un mot du segment présent dans la question est le complémentaire de $P(w_i)$:

$$\overline{P}(w_i) = 1 - P(w_i) \qquad [5]$$

Afin de capturer une réponse, représentée et délimitée par une Entité Réponse (ER) dans un segment, et depuis l'ensemble de ces probabilités :

$$P(\{w_i \in S\}) = \prod_{w_i \in Q \cap S} \overline{P}(w_i) \qquad [6]$$

Cependant comme cela a été envisagé, nous souhaitons prendre en compte un éloignement des mots de la question vis-à-vis de cette ER, et notamment considérer qu'une forte proximité des mots de la question autour d'une ER contribue à préférer cette dernière à une autre où l'information serait plus diffuse. Ce qui peut être envisagé d'une manière similaire à (Koehn *et al.*, 2003) au travers d'une mise à la puissance. En effet, le cas qui intéresse Koehn *et al.* concerne une probabilité de distorsion : l'ordre d'une expression et de sa traduction dans une autre langue peut être modifié. Cette probabilité de distorsion (pour prendre en compte l'alignement)

peut être estimée au travers d'un modèle de probabilité jointe, ou plus simplement par un paramètre approprié élevé à la puissance de la différence des positions de départ et de fin de cette expression traduite. Dans notre cas concret, l'élévation à la puissance peut permettre de tenir compte d'une fonction d'éloignement ou d'une distance, soit F_d cette fonction :

$$P(\{w_i \in S\}) = \prod_{w_i \in Q \cap S} \overline{P}(w_i)^{F_d} \qquad [7]$$

Soit *écart*, l'intervalle positionnel (et par conséquent la distance) entre une ER et le mot w_i ; et *nwQ*, le nombre de termes de la question compris dans cet intervalle positionnel. Il est possible d'envisager F_d comme un ratio entre *nwQ* et *écart*. En outre, l'impact de cette élévation à la puissance doit être moins important dans le cas ou le segment contient des termes de la question (il faut garder à l'esprit que la probabilité étant comprise entre 0 et 1, une élévation à une puissance supérieure diminue cette probabilité).

Pour les mêmes raisons que KSJ $-\log\frac{n_i}{N_S}$ est préféré à $\log\frac{N_S - n_i}{N_S}$ comme expression de $\overline{P}(w_i)$.

Aussi, en passant par le logarithme, le score final C défini en [7] d'un segment peut être calculé comme :

$$C(S) = \sum_{w_i \in Q \cap S} F_d \log\left(\overline{P}(w_i)\right) = \sum_{w_i \in Q \cap S} \frac{nwQ}{écart} \log\left(\frac{N_S}{n_i}\right) \qquad [8]$$

Ce dernier score conclut cette partie et constitue ce que nous appelons une *« compacité probabiliste »* et que nous expérimentons en section suivante par rapport à la *« compacité positionnelle »* présentée au début de cette section.

Remarque : comme indiqué au début de cette partie, nous avons déroulé le cheminement suivi pour obtenir, depuis un calcul de précision inspiré par la métrique *CWS*, un premier score de *« compacité positionnelle »* puis une extension de ce score vers une *« compacité probabiliste »*. Ce qui peut également être perçu depuis (en faisant correspondre les notations *nWQ* avec *Z* et *écart* avec *2R+1*) :

$$compacité(ERc_i) = \frac{1}{|QW|} \sum_{w \in QW} contribCompacité(w)_{ERc_i}$$

$$\approx \sum_{w \in QW} contribCompacité(w)_{ERc_i} = \sum_{w \in QW} \frac{|Z|}{2R+1} = \sum_{w \in QW} \frac{|nWQ|}{2R+1} = \sum_{w \in QW} \frac{|nWQ|}{écart}$$

$$C(S) = \sum_{w_i \in QW} contribCompacité(w_i)_{ERc_i} \times \log\left(\frac{N_S}{n_i}\right)$$

5. Résultats Expérimentaux

5.1. *Préalable : quels segments pour calculer l'IDF ?*

Jusqu'à présent nous avons éludé un problème qui se pose dans le calcul de la compacité probabiliste de la section précédente. En effet, elle nécessite de disposer des valeurs numériques correspondant à deux variables : n_i et N_s, respectivement le nombre de segments contenant un mot w_i et le nombre total de segments considérés. Cependant, comment doit-on définir ces segments ? Quelle est leur taille ? Voire leur justification théorique ?

En QR, nous avons vu qu'une Réponse était extraite depuis un document. Le corollaire de cela est que la zone dont la réponse est extraite constitue une sorte de justification (appelé aussi un support) pour cette réponse. En outre, lors des campagnes d'évaluation, un tel support doit souvent être produit et cela, avec une taille limitée (le plus souvent à 50 ou 250 octets). C'est aussi dans un segment d'une taille variable et ne correspondant pas forcément à celle d'une phrase (même si un tel découpage pourrait paraître plus naturel) que se trouvent une réponse et les mots de la question utilisés pour l'extraire.

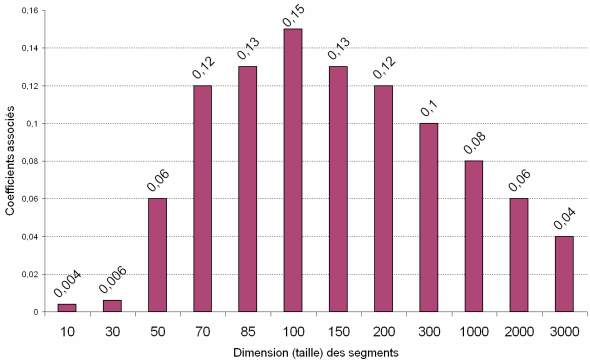
Aussi, et justement parce que nous n'avons *a priori* aucun moyen de connaître quelle sera exactement la taille du segment utilisé pour répondre à une question, nous faisons le choix de définir des segments de taille variable et de calculer une combinaison linéaire afin d'approximer la probabilité à la base de notre compacité :

$$P_m(w_i) = \sum_{s \in \{s_1, s_2, \ldots\}} \lambda_s P_s(w_i) \text{ avec } \sum_s \lambda_s = 1 \qquad [9]$$

Par ailleurs, afin de ne pas accorder une importance « surestimée » à la fréquence des mots, nous calculons ces segments sur des fenêtres de taille fixe et non glissantes (puisque sinon chaque mot serait compté plusieurs fois). Les différentes tailles et coefficients λ_s jouent un rôle de lissage. Ceux de nos expériences sont présentés sur le graphique 1. Nous les avons définis empiriquement et, avec un corpus de Questions/Réponses conséquent, ils pourraient faire l'objet d'un apprentissage. Les calculs d'IDF sont effectués sur l'intégralité du corpus de documents associé à la campagne EQueR.

5.2. *Cadre d'évaluation et résultats*

Ce que nous souhaitons évaluer et comparer est la capacité des deux compacités précédemment décrites à placer au moins, et à la meilleure des positions possibles, une réponse correcte parmi les 5 premières réponses autorisées pour chacune des questions factuelles considérées.



Graphique 1. *Dimensions et coefficients λ_s pour le calcul de $P_m(w)$.*

Ces questions proviennent de la tâche « Généraliste » de la campagne d'*É*valuation en *Que*stions-*R*éponses (EQueR, décrite dans Ayache, *et al.*, 2005). Nous utilisons les données et méthodologies de la campagne. Les questions à prendre en compte sont celles auxquelles notre système peut répondre. C'est-à-dire celles qui sont correctement étiquetées en type attendu, et pour lesquelles au moins une réponse correcte a été balisée, à l'aide d'un étiquetage en Entités Nommées, dans l'un des passages en amont de l'étape d'extraction des réponses. En effet, dans le cas contraire, et de par l'enchainement séquentiel des étapes présentées en figure 1, un silence ou une erreur fait que notre sQR ne peut extraire une réponse correcte. Or c'est précisément cette capacité à l'extraire au mieux que nous souhaitons étudier.

Enfin, nous nous intéressons uniquement aux *réponses courtes correctes*. 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 permet de l'établir. 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 de ces conditions n'est remplie.

Les évaluations présentées sont obtenues grâce à un script automatique utilisant une référence que nous avons construite à l'aide de l'ensemble des réponses des participants de la campagne EQueR. En effet, 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. Par ailleurs, sur les 19708 réponses courtes jugées manuellement par les évaluateurs de la campagne, notre référence obtient un taux d'accord de 97,5% avec ces jugements.

Résultats expérimentaux. Le tableau page suivante présente un bilan de nos expériences sur la compacité probabiliste et leur comparaison avec la compacité positionnelle. Les lignes sont numérotées de *a* à *i* pour faciliter leur commentaire et doivent être lues deux à deux : la première correspond à une évaluation *Stricte*, la seconde (marquée prime') à une évaluation tolérante (*Tol.*). Les colonnes correspondent à une ventilation des questions (factuelles) suivant leur objet : le nom d'une *Pers.*(onne), d'une *Org.*(anisation), une *Mesure* (telle qu'une valeur numérique, monétaire, une population, *etc.*), la *Date* d'un événement, le nom d'un *Lieu*, une *Manière* ou *Autres*. L'avant dernière colonne (%) contient les pourcentages de réponses correctes par rapport aux 400 questions factuelles de la campagne ; et, la ligne *a* en donne la ventilation suivant chacun des objets. Étant donné que les éventuels problèmes en amont de l'étape d'extraction (qui nous intéresse particulièrement dans cet article) influent sur les performances finales et par conséquent ces pourcentages, la dernière colonne (*Δ%)* présente les pourcentages de bonnes réponses par rapport aux maximums atteignables pour notre système. Lesquels sont présentés en lignes *b* et *b'*. Ainsi, au plus 241 (ou 250) réponses courtes correctes peuvent être trouvées dans les passages sur les 400.

L'expérience *c* est notre expérience de référence : pour extraire une réponse, elle utilise un calcul de compacité positionnelle basé sur les mots après l'utilisation d'une *Stoplist*. En cas d'occurrences multiples d'un mot, seule l'occurrence la plus proche de la réponse candidate est prise en compte. Il s'agit du calcul de compacité qui a été utilisé lors de nos participations aux campagnes de QR. Elle extrait de 40 à 44% des 400 réponses de la campagne ou de 67 à 71% des réponses envisageables.

Les expériences *d* à *i* utilisent le calcul de compacité probabiliste explicité en section 4 et l'approximation de l'IDF décrite en section 5.1. Les différences résident sur les unités utilisées : un calcul sur les mots (de *d* à *f*) ou sur les lemmes (de *g* à *i*). Par ailleurs, 3 autres distinctions, correspondant à 3 stratégies différentes, sont faites dans les cas d'occurrences multiples d'un même mot *(m)*, ou lemme *(l)* : *cM* si l'occurrence prise en compte est celle de la meilleure contribution à la compacité ; *cN* s'il s'agit de la plus proche ; ou *cTF* pour la somme de toutes les contributions de chacune des occurrences.

Il est possible de constater de toutes ces expériences que les performances des différentes compacités probabilistes envisagées *cM*, *cN*, *cTF*, qu'elles aient lieues sur les *mots* ou sur les *lemmes* sont à la fois comparables entre elles, et à celles de notre compacité positionnelle de référence. En outre, aucun des calculs ou des

stratégies envisagées n'apparaît dégager un profil apte à prendre en compte un objet particulier de question.

		Org.	Pers.	Mesure	Date	Lieu	Manière	Autres	TOTAL	%	Δ%
a		*28*	*95*	*77*	*42*	*65*	*26*	*67*	*400*	*100%*	
	Maximum envisageables pour notre système de Questions/Réponses										
b	*Stricte*	11	70	66	32	43	4	15	**241**	*60%*	*100%*
b'	*Tol.*	12	72	67	35	44	4	16	**250**	*63%*	*100%*
	***C(m)* : Compacité positionnelle (calculée sur les mots, occurrence la plus proche)**										
c	*Stricte*	8	48	41	23	26	4	11	**161**	*40%*	*67%*
c'	*Tol.*	8	52	44	24	32	4	13	**177**	*44%*	*71%*
	***cM(m)* : Compacité probabiliste (sur les mots, meilleure contribution en cas d'occurrences multiples)**										
d	*Stricte*	8	47	36	22	29	4	11	**157**	*39%*	*65%*
d'	*Tol.*	9	53	38	24	32	4	14	**174**	*44%*	*70%*
	***cN(m)* : Compacité probabiliste (sur les mots, occurrence la plus proche)**										
e	*Stricte*	8	46	36	22	29	4	11	**156**	*39%*	*65%*
e'	*Tol.*	9	53	38	24	32	4	14	**174**	*44%*	*70%*
	***cTF(m)* : Compacité probabiliste (sur les mots, somme de toutes les occurrences)**										
f	*Stricte*	8	50	35	22	29	4	12	**160**	*40%*	*66%*
f'	*Tol.*	9	55	38	23	32	4	14	**175**	*44%*	*70%*
	***cM(l)* : Compacité probabiliste (sur les lemmes, meilleure contribution en cas d'occurrences multiples)**										
g	*Stricte*	7	46	38	22	31	4	11	**159**	*39%*	*65%*
g'	*Tol.*	8	51	39	24	34	4	14	**174**	*44%*	*70%*
	***cN(l)* : Compacité probabiliste (sur les lemmes, occurrence la plus proche)**										
h	*Stricte*	7	44	38	22	31	4	11	**157**	*39%*	*65%*
h'	*Tol.*	8	50	39	24	34	4	14	**173**	*43%*	*69%*
	***cTF(l)* : Compacité probabiliste (sur les lemmes, somme de toutes les occurrences)**										
i	*Stricte*	7	49	37	22	32	3	12	**162**	*41%*	*67%*
i'	*Tol.*	8	53	38	25	35	4	14	**177**	*44%*	*71%*

Tableau 1. *Comparaison sur le nombre de réponses correctes pour différentes natures de questions et différents calculs de Compacité : positionnelle (lignes c et c') et probabiliste (lignes d à i'). Evaluations stricte ou tolérante.*

12. Conclusion et perspectives

Nous avons explicité le score de compacité positionnel mis en œuvre dans le processus d'extraction d'une réponse de notre système de Questions/Réponses. Puis, nous avons proposé et discuté une extension à un modèle probabiliste de ce score de compacité.

Nous avons modélisé ce score de compacité probabiliste en utilisant un critère d'importance des mots de la question au travers d'un critère IDF. Nous avons expérimenté différentes variations de ce score, et constaté qu'elles se comportent de manières comparables.

En outre, ce score nous apparaît intéressant car il ouvre des perspectives pour la prise en compte, directement au sein d'un calcul numérique, de certaines relations sémantiques telles que l'hyperonymie. En effet, puisqu'une notion d'importance est associée à chacun des mots de la question présent dans le voisinage d'une réponse candidate, un hyperonyme (qui apparaît alors comme un substitut) d'un de ces mots dans un passage se voit attribuer une importance moindre que celle qu'aurait eu ce même mot, à cette même place.

12. 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.

Koehn P., Och F.J., Marcu D., "Statistical Phrase-Based Translation", *Actes de " The Human Language Technology Conference 2003 (HLT-NAACL 2003)"*, Edmonton, Canada, 27 mai - 1ier juin 2003, p. 48 – 54.

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

Gillard L., Bellot P., El-Bèze M., « Influence de mesures de densité pour la recherche de passages et l'extraction de réponses dans un système de questions-réponses », *Actes de la 3ième Conférence en Recherche d'Informations et Applications (CORIA)*, Lyon, France, 15-17 mars 2006, p. 193-204.

Gillard L., Sitbon L., Blaudez E., Bellot P., El-Bèze M., "The LIA at QA@CLEF2006", *Dans les "Working Notes of the Cross Language Evaluation Forum (CLEF) 2006"*, Alicante, Espagne, 20-22 septembre 2006.

Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus V., Morarescu P. "FALCON - Boosting Knowledge for Answer Engines", *Actes de "The 9th Text REtrieval Conference"*, Gaithesburg, Maryland, USA, 16 au 19 novembre 2000, p. 479-488.

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

Radev D., Fan W., Qi H., Wu H. and Grewal A., "Probabilistic Question Answering on the Web", *Actes de "The 11th International World Wide Web Conference"*, Honolulu, Hawaii, USA, 7 au 11 mai 2002. p. 408-419. (http://www2002.org/presentations/fan.pdf).

Spärck Jones K., Walker S., Robertson S.E., "A probabilistic model of information and retrieval: development and status", Technical Reports 446, août 1998, Cambridge University Computer Laboratory. (http://www.cl.cam.ac.uk/TechReports/UCAM-CL-TR-446.html).

Subbotin M., Subbotin S. "Patterns of Potential Answer Expressions as Clues to the Right Answers", *Actes de "The 10th Text REtrieval Conference"*, Gaithesburg, Maryland, USA, 2001. p. 293-303.

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.

Voorhees E.M., Harman, D., *TREC Experiment and Evaluation in Information Retrieval*. MIT Press, chapter 10. p. 233-257, 2005.