TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Ajout d'informations contextuelles pour la recherche de passages au sein de Wikipédia

Romain Deveaud Eric SanJuan Patrice Bellot
LIA - Université d'Avignon
339, chemin des Meinajariès Agroparc BP 91228
84 911 Avignon Cedex 9
{romain.deveaud,eric.sanjuan,patrice.bellot}@univ-avignon.fr

Résumé. La recherche de passages consiste à extraire uniquement des passages pertinents par rapport à une requête utilisateur plutôt qu'un ensemble de documents entiers. Cette récupération de passages est souvent handicapée par le manque d'informations complémentaires concernant le contexte de la recherche initiée par l'utilisateur. Des études montrent que l'ajout d'informations contextuelles par l'utilisateur peut améliorer les performances des systèmes de recherche de passages. Nous confirmons ces observations dans cet article, et nous introduisons également une méthode d'enrichissement de la requête à partir d'informations contextuelles issues de documents encyclopédiques. Nous menons des expérimentations en utilisant la collection et les méthodes d'évaluation proposées par la campagne INEX. Les résultats obtenus montrent que l'ajout d'informations contextuelles permet d'améliorer significativement les performances de notre système de recherche de passages. Nous observons également que notre approche automatique obtient les meilleurs résultats parmi les différentes approches que nous évaluons.

Abstract. Traditional Information Retrieval aims to present whole documents that are relevant to a user request. However, there is sometimes only one sentence that is relevant in the document. The purpose of Focused Information Retrieval is to find and extract relevant passages instead of entire documents. This retrieval task often lacks of complement concerning the context of the information need of the user. Studies show that the performance of focused retrieval systems are improved when user manually add contextual information. In this paper we confirm these observation, and we also introduce a query expansion approach using contextual information taken from encyclopedic documents. We use the INEX workshop collection and evaluation framework in our experiments. Results show that adding contextual information significantly improves the performance of our focused retrieval system. We also see that our automatic approach obtains the best results among the different approach we evaluate.

Mots-clés : Recherche de passages, enrichissement de requêtes, contexte, Wikipedia, INEX, entropie.

Keywords: Focused retrieval, query expansion, context, Wikipedia, INEX, entropy.

1 Introduction

Les approches traditionnelles de Recherche d'Information (RI) cherchent à retrouver le ou les documents les plus pertinents par rapport à une requête. Parfois, seule une petite partie d'un document est très pertinente mais celui-ci est mal classé et l'information n'est pas présentée à l'utilisateur. La recherche de passages consiste à rechercher uniquement ces passages pertinents, en laissant le soin au système de déterminer le meilleur niveau de granularité des documents tout en évitant le recouvrement entre différents passages.

La recherche de passages restreints, ou *Focused Information Retrieval* (Trotman *et al.*, 2010; Arvola *et al.*, 2011), ajoute une contrainte qui consiste à limiter la taille totale du texte présenté à l'utilisateur. Cette taille est fixée par des contraintes techniques telles que la taille d'un écran d'un téléphone portable. Les passages extraits et présentés à l'utilisateur doivent également être lisibles et compréhensibles (Harman & Over, 2002). Cette tâche s'approche ainsi de celle du résumé automatique puisqu'il s'agit de récupérer les passages les plus importants dans un corpus, plutôt que de récupérer l'ensemble des passages pertinents (Dang, 2005; Dang & Owczarzak, 2008; Harman & Over, 2002). Cependant, à la différence des résumés automatiques, il ne s'agit pas de générer un résumé mais bien d'extraire des passages des documents, ces passages pouvant être utilisés dans un processus de recherche d'information interactive. La campagne d'évaluation INEX a été créée dans le but de développer la recherche de passages et d'éléments dans des documents XML. Dans ce contexte, un cadre d'évaluation de recherche de passages est utilisé depuis l'apparition de la tâche *Focused* de la piste Ad Hoc (Kamps *et al.*, 2008).

C'est lors de l'édition 2010 que la campagne INEX a vu apparaître une tâche de recherche de passages restreints. Un corpus de requêtes a été constitué en reprenant la méthodologie des précédentes campagnes, en retenant de façon prioritaire les requêtes dont les éléments de réponse étaient répartis dans plusieurs pages Wikipedia. Par ailleurs, chaque requête constituée d'un titre très court était complétée par une liste de termes (mots-clés ou syntagmes nominaux) précisant le contexte thématique de la recherche.

Au moyen d'un système de recherche de passages, nous proposons de montrer que non seulement ce contexte améliore les résultats de la recherche mais aussi qu'il est possible d'en retrouver une variante automatiquement qui aboutit à des hausses de performance comparables. Cela, dans le cadre d'un processus de recherche encyclopédique à partir de requêtes complexes, constitue une avancée notable à mettre en regard avec les résultats souvent décevants obtenus par les méthodes automatiques d'enrichissement.

2 Méthodologie

2.1 Extraction de phrases et attribution de scores

Dans toutes nos expérimentations nous utilisons le système que nous avons proposé comme étalon à la tâche de Question-Réponse d'INEX (SanJuan *et al.*, 2011). Ce système utilise Indri [1] pour l'indexation et l'extraction de documents ainsi qu'une approche par modèles de langue pour la RI (Metzler & Croft, 2004). Lors de l'indexation, les mots ne sont pas tronqués de manière à permettre des recherches exactes, et leurs positions dans les documents sont mémorisées. Ce système intègre également un étiqueteur morpho-syntaxique incluant l'étiquetage de la ponctuation : TreeTagger [2].

Nous utilisons une approximation rapide de l'algorithme LexRank (Erkan & Radev, 2004) guidée par la requête afin d'attribuer des scores à des phrases extraites du corpus. Ces phrases sont considérées comme des paquets de lemmes où seuls les adjectifs et les noms sont retenus. Un sous-graphe d'intersection est formé avec les phrases qui possèdent au moins un mot en commun avec requête et avec celles qui possèdent au moins un mot en commun avec ces premières phrases. Les phrases sont les sommets du sous-graphe, et une arête relie deux sommets si les deux phrases ont au moins un mot en commun. La valeur d'une arête reliant deux sommets est le nombre de mots en commun entre les deux phrases représentées par ces sommets. Chaque sommet reçoit un poids initial correspondant au nombre de mots de la requête présents dans la phrase, incrémenté d'une unité. L'algorithme LexRank attribue alors des scores à chacune des phrases du sous-graphe. Les 1 500 phrases possédant les scores les plus importants sont conservées.

1. www.lemurproject.org
2. www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

2.2 Recherche d'information ciblée par extraction de phrases

Les requêtes que nous utilisons proviennent de la collection fournie pour la tâche Ad Hoc de l'édition 2010 de la campagne INEX (Arvola *et al.*, 2011). Plusieurs types d'informations sont disponibles pour chaque *topic* (voir figure 1).

Figure 1 : Un extrait de *topic* de la tâche Ad Hoc d'INEX 2010.

```
<title>health risk coca leaf</title>
<phrasetitle>"health risk" "coca leaf" "traditional coca leaf consumption" "health study"</phrasetitle>
```

Le champ `<title>` contient une requête utilisateur de quelques mots-clés, telle qu'elle serait entrée dans un moteur de recherche grand public. Le champ `<phrasetitle>` contient quant à lui des syntagmes nominaux et des mots-clés explicitement ajoutés par l'utilisateur pour préciser le contexte de sa recherche et ainsi aider le système à retrouver les passages susceptibles de l'intéresser. Les expérimentations menées par (Vechtomova, 2005) montrent notamment qu'un ajout manuel, par l'utilisateur, de multi-mots liés au contexte apporte une amélioration des résultats.

Recherche de mots au sein de séquences Dans notre première approche, les mots-clés et syntagmes nominaux supplémentaires de la requête sont ajoutés au modèle de langage comme des séquences de mots autorisant jusqu'à 4 insertions dans le cas de multi-mots. On voit par exemple dans l'extrait de passage ci-dessous que le système a autorisé l'insertion des deux mots « or » et « cuca » au sein du multi-mot « coca leaf », tout en respectant l'ordre d'apparition précisé par l'utilisateur dans le champ `<phrasetitle>` :

> *He inquires about the **coca** or cuca **leaf** from Peru ...*

Nous considérons également le cas de fenêtres non ordonnées. Toujours pour le même *topic*, notre système a extrait le passage suivant en autorisant des insertions et l'inversion de l'ordre de mots :

> *... purchasing , personnel , **risk** management , environmental **health** and safety ...*

Cette approche, qui ne considère donc que la position relative des mots sans utilisation de pondération, s'est révélée efficace lors de précédentes participations à INEX (SanJuan & Ibekwe-SanJuan, 2010).

Pondération des mots selon leur importance contextuelle La deuxième approche que nous expérimentons consiste à ajouter des mots ou des multi-mots pondérés selon leur importance informative par rapport au contexte de la recherche. En effet, dans un document, certains mots sont plus importants que d'autres en terme de valeur informative. Dans cette approche nous utilisons une mesure d'entropie, basée notamment sur la fréquence des mots dans le document, reconnue pour sa capacité de mise en évidence des mots ou des multi-mots les plus informatifs. Ces mots sont directement extraits d'une page Wikipedia sélectionnée automatiquement à partir de la requête utilisateur. C'est donc dans cette page que le contexte de recherche va pouvoir être exploré comme nous le détaillons dans la section 3.

Extraction de phrases La partie du système permettant de récupérer les passages est commune à ces différentes approches. Nous procédons d'abord à une extraction des 50 documents les plus pertinents en utilisant une approche par modèles de langue, où les probabilités sont estimées par maximum de vraisemblance et lissées par une règle de Dirichlet. Ce lissage permet d'éviter les probabilités nulles ; il est particulièrement adapté aux requêtes formées de mots-clés (Zhai & Lafferty, 2004). Le texte des documents extraits lors de cette première passe sont concaténés et segmentés en phrases. Les documents de la collection contiennent de nombreux tableaux et listes, mais ces informations ne peuvent pas être intégrées dans des passages, c'est pourquoi leur contenu est écarté. Les phrases sont ordonnées par score décroissant. Chaque phrase est alors considérée comme une suite de mots autorisant l'insertion toutes les insertions de caractères n'étant pas des lettres ni des chiffres. Tous les passages qui vérifient ces conditions sont extraits des 50 documents précédemment retenus. Les scores des phrases obtenus par la méthode détaillée dans la section 2.1 sont utilisés pour pondérer les passages. Ils sont enfin ordonnés par pondération décroissante. Voici un exemple de passage extrait par le système qui répond au *topic* présenté dans la figure 1 :

> *Because cocaine is a stimulant , a user will often drink large amounts of alcohol during and after usage or smoke cannabis to dull "crash" or "come down" effects and hasten slumber.*

3 Recherche automatique du contexte

Dans les approches que nous avons présentées, nous enrichissons la recherche de passages pertinents avec des éléments issus du contexte de la requête. Nous avons vu que ce contexte pouvait notamment être représenté par plusieurs termes ajoutés manuellement par l'utilisateur. Plusieurs études ont exploré l'utilisation de Wikipedia comme collection externe pour l'enrichissement de requêtes (Koolen *et al.*, 2009; Li *et al.*, 2007; Milne *et al.*, 2007; Xu *et al.*, 2009). Seulement, ces études ne rapportent des résultats que pour la recherche de documents entiers et non pour la recherche de passages.

Nous avons mis en place une bibliothèque logicielle [3] permettant d'associer une page Wikipédia à une requête. Nous interrogeons le moteur de recherche de Wikipédia et nous sélectionnons le premier résultat renvoyé comme la page contenant les informations contextuelles. Ceci nous permet notamment d'éviter les pages de désambiguïsation qui ne contiennent que des liens vers d'autres pages. Les interrogations sont effectuées sur la version en ligne de l'encyclopédie ce qui nous assure d'avoir les informations les plus à jour. Par exemple, une requête « bonaparte empereur » nous permettra de récupérer automatiquement la page Wikipedia de Napoleon I^{er}.

Une fois la page récupérée, le texte brut est extrait. Une mesure d'entropie est calculée pour les tous les n-grammes présents dans le texte. Soit une suite de mots $S = (w_1, ..., w_n)$, la mesure d'entropie H est calculée selon la formule suivante :

$$H(S) = -\sum_{i=1}^{n} P_{Wiki}(w_i) \times \log_2(P_{Wiki}(w_i))$$

où les probabilités d'apparition des mots sont calculées sur l'ensemble des mots de la page Wikipedia associée à la requête. Cette mesure permet de refléter l'importance relative de tous ces mots dans la page Wikipedia. En effet dans le cas de la page de Napoleon I^{er}, le terme « premier empereur » a bien plus d'importance que « nombreux ouvrages » dans le contexte d'une recherche sur Napoleon Bonaparte.

Ces scores d'entropie nous permettent ainsi de pondérer différents mots selon leur importance informative dans le contexte de la requête initiale de l'utilisateur. L'utilisation d'une mesure d'entropie pour l'extraction et la pondération de mots a déjà prouvé son efficacité dans le cadre d'une recherche de livres entiers (Deveaud *et al.*, 2011), c'est pourquoi nous l'intégrons à notre approche de recherche de passages décrite dans la section 2.2. Dans nos expérimentations, nous n'extrayons pour l'instant que des unigrammes des pages Wikipedia ; si on reprend la notation ci-dessus, $S = (w_1)$.

4 Expérimentations et résultats

Nos expérimentations sont menées avec les 107 *topics* de la tâche Ad Hoc d'INEX 2010 et leurs jugements de pertinence. Nous comparons notre approche de récupération automatique du contexte avec les approches manuelles que nous avons décrites dans la section 2.2 ainsi qu'avec les résultats du système de l'Indian Statistical Institute (ISI2010) qui a obtenu les meilleurs résultats de recherche de passages restreints d'INEX 2010. Ce système utilise des méthodes de classification directement sur les éléments XML présents dans les documents afin d'extraire des éléments entiers, et non pas seulement des phrases. La *baseline* que nous proposons dans cet article utilise des requêtes utilisateur uniquement composées de mots-clés (le champ `<title>` des *topics*).

Les résultats obtenus pour l'extraction de passages de 1 000 caractères au maximum sont présentés dans le tableau 1. Nous présentons également dans le tableau 2 les résultats obtenus dans le cas d'une recherche de passages où le nombre de caractères n'est pas limité. Lors d'une recherche de passages, les utilisateurs vont être attentifs aux tout premiers résultats renvoyés par le système, c'est pourquoi nous privilégions une mesure à 1% de rappel.

L'approche d'ajout automatique d'informations contextuelles est celle qui fonctionne le mieux pour les deux types d'extraction de passages. L'amélioration observée pour la recherche de passages restreints par rapport au système ISI2010 est de l'ordre de 5%. On observe dans le tableau 2 que les précisions contextuelles apportées par les utilisateurs améliorent les performances de façon significative par rapport à la *baseline*, ce qui confirme les constatations de (Vechtomova, 2005). L'utilisation d'une fenêtre ordonnée pour la recherche de multi-termes semble mieux marcher dans le cas où les passages sont limités à 1000 caractères. En effet, l'ordre des mots défini

3. `mirimiri.org`

TABLE 1: Résultat de recherche de passages restreints à 1 000 caractères, en terme de précision interpolée à différents niveaux de rappel (iP).

Approche	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]
contexte automatique	0,565	0,167	0,080	0,020
ISI2010	–	0,153	0,019	0,000
contexte utilisateur / fenêtre ordonnée	0,492	0,148	0,000	0,000
contexte utilisateur / fenêtre non ordonnée	0,529	0,147	0,000	0,000
baseline	0,486	0,141	0,000	0,000

TABLE 2: Résultat de recherche de passages, en terme de précision interpolée à différents niveaux de rappel (iP). T-test unilatéral apparié sur la significativité de l'amélioration des différentes méthodes vis-à-vis de la *baseline* sur les *topics* avec iP[0.00] < 1 (* : p < 0,1 ; ** : p < 0,05) .

Approche	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]
contexte automatique	$0,647^{**}$	$0,565^{*}$	$0,453^{**}$	$0,358^{**}$
contexte utilisateur / fenêtre non ordonnée	$0,634^{**}$	$0,533^{**}$	$0,385^{**}$	$0,279^{*}$
contexte utilisateur / fenêtre ordonnée	$0,613^{*}$	$0,522^{**}$	$0,385^{**}$	$0,288^{*}$
baseline	0,585	0,504	0,359	0,265

par l'utilisateur prend plus d'importance dans le cas où le système doit chercher des passages courts. *A contrario*, le système est plus performant avec des fenêtres non ordonnées lorsqu'il n'y a pas de limitation dans la taille des passages.

De son côté, l'approche par détection automatique du contexte et extraction des mots informatifs par mesure d'entropie obtient les meilleurs résultats pour les mesures présentées. Ces améliorations peuvent êtres expliquées par le fait que le vocabulaire introduit par les utilisateurs est assez redondant, ce qui handicape l'extraction de nouvelles phrases. Inversement, les mots extraits des pages Wikipedia associés aux requêtes apportent du vocabulaire nouveau. C'est cette extension du vocabulaire à des mots rares et spécifiques permet au système d'extraire des phrases pertinentes qui n'auraient pas été présentes autrement. En effet le mot « cocaine » n'était pas mentionné dans le *topic* présenté dans la section 2.2, et il a été désigné par le système comme un mot important compte tenu du contexte de la recherche, ce qui a permis de récupérer le passage suivant :

> *The production , the distribution and the sale of* ***cocaine*** *products is restricted (and illegal in most contexts) in most countries.*

Parfois, certains mots extraits des pages Wikipedia ne sont pas importants dans le contexte de la recherche de l'utilisateur. La pondération apportée par la mesure d'entropie permet alors de refléter leur importance contextuelle et de limiter leur effet négatif.

5 Conclusion

Nous avons présenté dans cet article un système de recherche de passages restreints. Cette approche consiste à former des passages pertinents à partir de phrases extraites des documents, et à les renvoyer à l'utilisateur. Nous avons présenté le système d'extraction de phrases que nous utilisons qui sert également de système étalon pour la tâche Question-Réponse de la campagne d'évaluation INEX.

Nous avons proposé une méthode automatique d'enrichissement de requêtes avec des mots fortement liés au contexte de la recherche. Ces mots sont extraits automatiquement des pages Wikipedia associées aux requêtes et pondérés à l'aide d'une mesure d'entropie. Nous utilisons cette mesure pour pondérer les mots extraits afin de refléter leur importance contextuelle au sein de la page Wikipedia.

Ces recherches ont bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR 2010 CORD 001 02) en faveur du projet CAAS.

Nous avons comparé cette méthode automatique à une méthode manuelle où le contexte est précisé par l'utilisateur au moment de la requête. Nous avons également reporté les scores obtenus par le meilleur système d'INEX 2010. L'approche automatique que nous proposons est la méthode qui obtient les meilleurs résultats pour les différentes mesures proposées, pour des passages libres ou restreints à 1 000 caractères.

Références

ARVOLA P., GEVA S., KAMPS J., SCHENKEL R., TROTMAN A. & VAINIO J. (2011). Overview of the inex 2010 ad hoc track. In S. GEVA, J. KAMPS, R. SCHENKEL & A. TROTMAN, Eds., *Comparative Evaluation of Focused Retrieval*, Lecture Notes in Computer Science. Springer Berlin / Heidelberg.

DANG H. T. (2005). Overview of DUC 2005. In *Proceedings of the 2005 Document Understanding Workshop*.

DANG H. T. & OWCZARZAK K. (2008). Overview of the tac 2008 update summarization task. In *Text Analysis Conference (TAC)*.

DEVEAUD R., BOUDIN F. & BELLOT P. (2011). Lia at inex 2010 book track. In S. GEVA, J. KAMPS, R. SCHENKEL & A. TROTMAN, Eds., *Comparative Evaluation of Focused Retrieval*, Lecture Notes in Computer Science : Springer Berlin / Heidelberg.

ERKAN G. & RADEV D. R. (2004). Lexrank : graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, **22**, 457–479.

HARMAN D. & OVER P. (2002). The duc summarization evaluations. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, p. 44–51, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

KAMPS J., PEHCEVSKI J., KAZAI G., LALMAS M. & ROBERTSON S. (2008). Inex 2007 evaluation measures. In N. FUHR, J. KAMPS, M. LALMAS & A. TROTMAN, Eds., *Focused Access to XML Documents*, volume 4862 of *Lecture Notes in Computer Science*, p. 24–33. Springer Berlin / Heidelberg.

KOOLEN M., KAZAI G. & CRASWELL N. (2009). Wikipedia pages as entry points for book search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, p. 44–53, New York, NY, USA : ACM.

LI Y., LUK W. P. R., HO K. S. E. & CHUNG F. L. K. (2007). Improving weak ad-hoc queries using wikipedia as external corpus. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, p. 797–798, New York, NY, USA : ACM.

METZLER D. & CROFT W. B. (2004). Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, **40**, 735–750.

MILNE D. N., WITTEN I. H. & NICHOLS D. M. (2007). A knowledge-based search engine powered by wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, p. 445–454, New York, NY, USA : ACM.

SANJUAN E., BELLOT P., MORICEAU V. & TANNIER X. (2011). Overview of the 2010 qa track. In S. GEVA, J. KAMPS, R. SCHENKEL & A. TROTMAN, Eds., *Comparative Evaluation of Focused Retrieval*, Lecture Notes in Computer Science : Springer Berlin / Heidelberg.

SANJUAN E. & IBEKWE-SANJUAN F. (2010). Multi word term queries for focused information retrieval. In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, p. 590–601. Springer Berlin / Heidelberg.

TROTMAN A., JIA X.-F. & GEVA S. (2010). Fast and effective focused retrieval. In S. GEVA, J. KAMPS & A. TROTMAN, Eds., *Focused Retrieval and Evaluation*, volume 6203 of *Lecture Notes in Computer Science*, p. 229–241. Springer Berlin / Heidelberg.

VECHTOMOVA O. (2005). The role of multi-word units in interactive information retrieval. In D. LOSADA & J. FERNÁNDEZ-LUNA, Eds., *Advances in Information Retrieval*, volume 3408 of *Lecture Notes in Computer Science*, p. 403–420. Springer Berlin / Heidelberg.

XU Y., JONES G. J. & WANG B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, p. 59–66, New York, NY, USA : ACM.

ZHAI C. & LAFFERTY J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, **22**, 179–214.