Recherche d'information et traitement automatique des langues : collaboration, synergie et convergence

Vincent Claveau, Jian-Yun Nie

Recherche d'information et traitement automatique des langues : collaboration, synergie et convergence

Introduction au numéro spécial sur la recherche d'information et le traitement automatique des langues

Vincent Claveau* — **Jian-Yun Nie****

* *IRISA-CNRS*
Campus de Beaulieu, Rennes, France
vincent.claveau@irisa.fr
** *Dépt. Informatique et Recherche Opérationnelle, Université de Montréal*
Montréal, QC, Canada
nie@IRO.umontreal.CA

1. Introduction

La recherche d'information (RI) et le traitement automatique des langues (TAL) ont beaucoup en commun : ils traitent tous deux de langues naturelles et de textes. Il serait donc naturel qu'il y ait beaucoup d'interactions entre les communautés RI et TAL. Pourtant, à quelques exceptions notables près, elles ont longtemps évolué indépendamment, développant chacune un corpus de techniques et de connaissances propres, sans beaucoup de contacts fructueux.

Du point de vue de la RI, les concepts et outils du TAL sont considérés avec défiance, comme le traduit la citation célèbre de Spärck-Jones (1999) :
« *It is not clear, [either] that NLP is required for some tasks that are closely related to ordinary retrieval.* »
En plus de leur éventuelle inutilité, les processus de TAL sont souvent vus comme trop

TAL. Volume 56 - n°3/2015, pages 7 à 22

coûteux pour être appliqués aux gigantesques ensembles de textes[1] qu'affectionnent les chercheurs en RI.

Du point de vue du TAL, les outils et concepts de la RI sont relativement mal connus, du moins pour ce qui est de leurs développements récents. On réduit souvent les techniques de RI en la pondération TF-IDF. Ainsi, dans les actes de la conférence TALN de 2007 à 2011, la pondération TF-IDF était utilisée dans trente et un articles, souvent comme étant « état de l'art », alors que la pondération plus avancée Okapi-BM25 (Robertson *et al.*, 1998), véritable standard de la RI, ne l'était que dans quatre articles. Cet état de fait met bien en évidence le manque du rapprochement entre les deux communautés de chercheurs.

Ce rapprochement des deux communautés nous semble cependant inexorable. Il est le fait, d'une part, d'une évolution des besoins : à l'ère du « big data », il faut savoir à la fois gérer la quantité et faire sens des textes. D'autre part, la convergence des techniques des sciences des données, notamment l'apprentissage artificiel et les approches numériques, offre un socle commun en terme d'outillage et de culture.

Ce numéro spécial de la revue TAL propose de faire le point sur cette convergence du TAL et de la RI, en présentant trois articles à la jonction des deux domaines. Ces derniers ont été sélectionnés, parmi douze articles soumis, par le comité de rédaction de la revue complété par un comité spécifique à ce numéro dont la liste est donnée en section 6, réunissant ainsi experts du TAL et de la RI. Nous tenons à remercier les membres de ces deux comités, qui, chacun avec des éclairages propres à leur domaine, ont permis d'apporter à tous les articles des retours pertinents et constructifs.

Le premier article de ce numéro, « Analyse en dépendance et classification de requêtes en langue naturelle, application à la recommandation de livres » d'Anaïs Ollagnier, Sébastien Fournier et Patrice Bellot, s'intéresse à la recherche de livres à l'aide de requête en langage naturel. Pour cette tâche, proposée dans le cadre de la campagne d'évaluation INEX (Bellot *et al.*, 2014), les auteurs proposent l'utilisation d'outils du TAL (classification de textes, détection de relations sémantiques à partir d'analyses syntaxiques en dépendance) afin d'améliorer les performances de leur système de recherche sur des requêtes complexes.

Le second article, « Recherche d'information sémantique : état des lieux » d'Haïfa Zargayouna, Catherine Roussey et Jean-Pierre Chevallet, propose un état de l'art en RI sémantique. Ce sous-domaine de la RI cherche à représenter l'information sous forme conceptuelle, pour s'abstraire, d'une part, de la représentation sous forme de mots, et pour permettre, d'autre part, l'utilisation d'ontologies, de forme de raisonnement... Évidemment, le passage du texte aux concepts englobe plusieurs problématiques propres au TAL (annotation sémantique automatique, inférence textuelle...). L'article présente ainsi un panorama des ressources et des techniques utilisées dans ce cadre, puis leur utilisation combinée dans les systèmes de RI.

1. Par exemple, la collection ClueWeb contient plus d'un milliard de pages Web dans dix langues différentes.

Le dernier article, « Recherche d'information précise dans des sources d'information structurées et non structurées : défis, approches et hybridation » de Brigitte Grau, Anne-Laure Ligozat et Martin Gleize, s'intéresse aux applications de questions-réponses (ou QA pour l'anglais *Question-Answering*). Il présente d'une part un panorama des différentes approches existantes dans ce domaine, en distinguant celles fondées sur la recherche dans des textes et celles fondées sur la recherche dans des bases de connaissances. Les principales techniques développées sont décrites et comparées. D'autre part, l'article propose plusieurs pistes pour le futur des systèmes de QA, et notamment une approche hybride qui combine les recherches dans les textes et dans les bases de connaissances.

Dans le reste de cette préface, nous replaçons ces différentes contributions dans le spectre plus large des interactions entre TAL et RI. Dans la section suivante, nous présentons quelques concepts issus de travaux en RI qui sont utilisés dans des tâches de TAL, et à l'inverse, la section 3 synthétise les apports du TAL dans les tâches de RI. Au-delà de ces travaux, la section 4 présente des applications dans lesquelles RI et TAL sont plus finement intriqués, cherchant à offrir dans le même temps l'accès à l'information et la gestion des phénomènes de la langue.

2. Concepts de RI dans des tâches du TAL

La représentation des documents au sein d'un système de RI et le calcul de similarité entre ces représentations sont deux problèmes centraux de la RI. Les nombreuses techniques développées pour y répondre ont été évidemment exploitées au-delà de la RI, notamment dans certaines tâches du TAL.

2.1. *Représentation*

Longtemps tournée vers la recherche dans des grandes bases de documents, la RI a développé différentes approches pour offrir une représentation des textes répondant à certaines contraintes. Elle doit notamment : être compacte, mais néanmoins décrire au mieux le contenu du texte ; être calculable avec une faible complexité temporelle et mémoire ; ne nécessiter aucune ressource externe, ou des ressources facilement disponibles ; permettre un calcul de similarité rapide et adapté à la tâche.

À ce titre, la représentation la plus communément employée en RI est celle dite en sac de mots : les documents sont considérés comme un multi-ensemble de mots-formes (formes graphiques), non structuré, sans information sur la séquentialité des mots dans le texte.

Usuellement, on calcule pour chaque mot présent dans le document une valeur reflétant son importance comme descripteur du document. Cette étape de pondération a généré un très grand nombre de travaux dans la communauté RI ; de nombreuses formules de pondération ont été proposées, étudiées et comparées. La plus célèbre d'entre elles est le TF-IDF ; bien que proposée dans les années 70, elle reste, souvent

à tort, celle utilisée majoritairement dans les travaux du TAL comme nous l'évoquions dans l'introduction.

Enfin, ces mots et leur valeur associée sont utilisés pour représenter le document dans son ensemble. Là encore, plusieurs modélisations ont été proposées, comme les bien connus modèle de langues (Ponte et Croft, 1998) et modèle vectoriel (Salton et McGill, 1983). Ce dernier est largement employé en RI mais également dans beaucoup d'applications du TAL où il sert à représenter les textes ou des segments de textes. Cette représentation vectorielle est plus particulièrement utilisée en entrée de classifieurs (SVM, forêts aléatoires, bayésien naïf, réseaux de neurones) ; elle est donc au cœur des travaux s'interprétant comme une tâche de classification (détection de spam, assignation de polarité ou de sentiment, attribution d'auteur...). Elle est d'ailleurs employée pour l'étape de classification de requêtes dans l'article de Ollagnier *et al.* (2015).

2.2. *Classification et RI*

Au-delà de la représentation vectorielle vue précédemment, les liens entre la RI et la classification sont plus profonds. En effet, la recherche de documents à partir d'une requête peut être vue comme une tâche de classification, et les moteurs de recherche comme des classifieurs : étant donné une requête, le moteur doit classer les documents entre pertinents et non pertinents (ou les classer du plus pertinent au moins pertinent), et il le fait sur la base de la proximité entre la représentation du document et celle de la requête. Grâce à cette capacité à calculer des proximités entre des textes, les moteurs de recherche peuvent par ailleurs être utilisés directement comme des classifieurs de types k-plus-proches voisins. Ils ont donc été logiquement employés comme tels dans différentes tâches pouvant être mises en œuvre par de la classification de documents (analyse de sentiment, recommandation par le contenu, détection de spam...).

Il faut également noter que cette utilisation des systèmes de RI comme classifieurs peut porter sur d'autres choses que sur des documents bien formés. Dans plusieurs travaux, le texte considéré n'est pas forcément un document. Par exemple, dans (Claveau *et al.*, 2014), il est montré que les moteurs de recherche peuvent être utilisés pour comparer des ensembles de contextes de mots dans le but de construire des lexiques distributionnels. En segmentation thématique, des systèmes de RI peuvent aussi servir à détecter les ruptures thématiques (Hearst, 1997 ; Claveau et Lefèvre, 2015) entre des portions successives d'un flux de textes.

2.3. *La RI pour évaluer le TAL*

Un des apports de la RI pour le TAL, indirect mais important, concerne l'évaluation. La communauté RI a très tôt commencé à formaliser des cadres d'évaluation quantitative (dont les mesures d'évaluation, comme le rappel, la précision, mais aussi les tests de significativité statistique), développer des collections de données, organiser

des compétitions[2]. L'idée centrale pour une évaluation objective est que les différents systèmes et méthodes doivent se comparer sur les mêmes données, et vis-à-vis d'une vérité terrain (*ground truth*). Pour permettre la reproductibilité et la comparabilité, la communauté RI a par ailleurs privilégié l'accessibilité des données, ouvertes ou avec un coût d'acquisition faible, et la mise à disposition des scripts d'évaluation.

Outre l'exemple de ces bonnes pratiques en matière d'évaluation quantitative, le cadre de la recherche documentaire a aussi largement été exploité pour servir à l'évaluation indirecte, par la tâche, de différents processus de TAL. Il s'agit dans ce cas d'étudier l'impact d'un outil du TAL sur un moteur de recherche en comparant les résultats avec et sans l'outil de TAL, ou de comparer les résultats obtenus avec plusieurs de ces outils.

3. Concepts de TAL dans des tâches de RI

L'importance de bien prendre en compte la complexité de la langue semble une évidence pour les tâches relevant de la RI : si le contenu langagier d'un document ou d'une requête est bien compris, les performances de la recherche ne peuvent être que meilleures. Deux phénomènes touchent particulièrement les systèmes de RI : l'**ambiguïté** (deux énoncés identiques ont, en contexte, des sens différents) et le **paraphrasage** (deux énoncés différents, en surface, ont un même sens). Le premier va ainsi avoir tendance à dégrader la précision, et le second le rappel du système. Pourtant, comme nous l'évoquions précédemment, les systèmes de RI classiques représentent le texte de manière très pauvre avec les sacs de mots, et calculent des similarités sur la base de comptages au sein de ces sacs de mots. Plusieurs travaux ont donc exploré les effets d'une meilleure prise en compte des phénomènes connus de la langue, en appliquant des méthodes ou des outils usuels du TAL. Un état de l'art encore d'actualité peut être trouvé dans (Moreau et Sébillot, 2005) ; nous ne présentons dans cette section que les grandes problématiques abordées par ces travaux en les distinguant selon les phénomènes considérés : morphologiques, syntaxiques ou sémantiques.

3.1. *Morphologie*

Les problèmes posés par la variation morphologique dans les systèmes de RI ont été notés très tôt. En effet, dans la représentation sac de mots, deux formes graphiques différentes, même légèrement (marque du pluriel par exemple), vont correspondre à deux entrées différentes. Pour autant, ces variations de forme marquent des différences sémantiques qui ne sont pas toujours pertinentes pour la tâche de RI ; il s'agit donc d'un problème relevant du paraphrasage. Par exemple, un utilisateur cherchant des documents avec une requête '*dog*' pourra accepter des documents contenant '*dogs*'. Ces variations ont pour conséquence qu'une requête contenant une forme ne sera pas

2. La première compétition TREC a été organisée en 1992.

appariée à un document ne contenant que l'autre forme. Et plus généralement, les pondérations de ces formes au sein d'un document, fondées sur des comptages, seront sous-évaluées.

Pour traiter ce problème, des outils de racinisation (*stemming*) ont été développés en RI opérant par conflation, c'est-à-dire regroupant les variantes sous une forme standardisée qui sera celle gardée dans le sac de mots. Ces outils diffèrent des techniques de lemmatisation issues du TAL, à la fois dans leur fonctionnement, et dans leur couverture des variations morphologiques traitées. Les raciniseurs sont le plus souvent des algorithmes reposant sur des heuristiques (suppression des préfixes et suffixes connus, suppression des diacritiques...) et traitent non seulement les cas de flexion, mais aussi certains cas de dérivation (Lovins, 1968 ; Porter, 1980 ; Savoy, 1993). La comparaison de la racinisation avec des outils de lemmatisation montre des résultats variables selon les langues et les tâches de RI (Moreau *et al.*, 2007 ; Savoy, 2002).

Si la racinisation est très largement utilisée en RI, la prise en compte d'autres phénomènes morphologiques est plus rare. Certains chercheurs ont néanmoins proposé des approches permettant de traiter plus finement les cas de dérivation (et aussi de flexion), dans un grand nombre de langues, et sans besoin de connaissances ou de ressources externes, en s'appuyant, par exemple, sur des analogies formelles (Gaussier, 1999 ; Moreau *et al.*, 2007).

Dans certains domaines de spécialité, les phénomènes morphologiques peuvent être plus riches et sont autant de freins à l'accès à l'information. Ainsi, dans le domaine biomédical, les cas de composition sont fréquents dans la construction des termes spécialisés. Leur prise en compte permet ainsi de noter que `stomachalgie`, `gastrodynie` et `maux d'estomac` sont synonymes, ou de mettre en correspondance une requête sur le `foie` avec un document portant sur l'`hépatite` (Claveau et Kijak, 2013). Notons enfin que certaines langues nécessitent également des modules morphologiques dédiés traitant par exemple de l'agglutination (allemand, turc...) (Haddad et Bechikh Ali, 2014), de la voyellation (par exemple pour la RI arabe) (Grefenstette *et al.*, 2005) et de segmentation (chinois, japonais) (Peng *et al.*, 2002).

3.2. *Syntaxe*

Avec la représentation sac de mots, la construction syntaxique des textes est complètement ignorée. D'une part, il n'est plus possible de distinguer « `La victoire de la France sur l'Italie` » de « `la victoire de l'Italie sur la France` » ou « `une pomme est tombée à terre` » de « `une pomme de terre est tombée` ». Il s'agit donc d'un problème d'ambiguïté de la représentation des textes, que plusieurs travaux ont bien entendu cherché à résoudre, ou amoindrir, en conservant des informations syntaxiques plus ou moins riches. D'autre part, des énoncés de même sens peuvent se trouver sous des formes différentes ; considérons par exemple (Strzalkowski *et al.*, 1999) les groupes de mots `information retrieval, retrieval of information, retrieve more information` et `information`

`that is retrieved`... Il s'agit dans ce cas d'un problème de paraphrasage, qui peut être résolu si l'on est capable de ramener ces formes à la relation `retrieve+information` où `retrieve` est l'élément tête et `information` son modifieur.

Selon les travaux, les phénomènes syntaxico-sémantiques considérés (et la façon de les nommer) varient. Il peut s'agir d'expressions multi-mots, de syntagmes, de phrasèmes, de mots composés, de collocations, etc., avec des propriétés également variables : expressions plus ou moins figées, continues ou discontinues, non (ou faiblement) compositionnelles (`pomme de terre, greenhouse gas effect`), ou compositionnelles (`sténose aortique`). Ces expressions sont détectées avec des méthodes qui peuvent reposer sur des indices purement numériques (par exemple, calculs de Pointwise Mutual Information (Acosta *et al.*, 2011)), sur des analyses de surface (patrons de parties-du-discours), sur des analyses syntaxiques complètes, ou bien encore sur l'utilisation de ressources externes (par exemple des lexiques et/ou des outils d'annotation comme l'UMLS et l'analyseur MMTx (Aronson et Lang, 2010) pour le domaine biomédical (Shen et Nie, 2015), ou des logs de moteurs de recherche généralistes (Chapelle et Chang, 2011)).

Une fois détectés, ces liens syntaxiques peuvent être utilisés dans le système de RI de différentes façons. Les expressions multi-mots peuvent servir à étendre les requêtes ; c'est l'approche utilisée par Ollagnier *et al.* (2015) dans ce numéro. Elles peuvent aussi être considérées dès l'indexation de manière figée, comme étant un seul mot (par exemple `retrieve_information`). Cette solution est facile à mettre en œuvre puisqu'elle ne remet pas en cause l'architecture des systèmes de RI reposant sur la représentation sac de mots. Mais des travaux ont également montré qu'il était possible d'inclure une information syntaxique riche dans la représentation des documents, et donc dans le calcul de similarité (Maisonnasse *et al.*, 2008 ; Gao *et al.*, 2004), intégrant ainsi dans le même temps les problèmes de l'ambiguïté et du paraphrasage. Ces travaux demandent alors une connaissance fine des processus de RI mais permettent une représentation plus riche du contenu des documents.

3.3. *Sémantique et connaissances*

Bien entendu, une meilleure prise en compte de la sémantique a été envisagée très tôt en RI. Dans une certaine mesure, celle-ci a visé à résoudre les problèmes d'ambiguïté, par exemple en identifiant le sens précis d'un mot-forme (Zhong et Ng, 2012), mais la majorité des travaux se sont intéressés au problème du paraphrasage. Dans ce dernier cas, les outils ou ressources utilisées ont pour but d'enrichir la description sac de mots avec des énoncés équivalents pour faciliter l'appariement entre la requête et les documents pertinents. Comme précédemment, les deux questions qui se posent sont celle de l'obtention de ces informations sémantiques et celle de leur intégration dans le système de RI.

Les lexiques sémantiques externes ont largement été utilisés. Bien que les premières expériences rapportaient des résultats négatifs[3], ces ressources apportent des gains parfois importants si elles sont adaptées aux documents et si elles sont bien intégrées au calcul de similarité. Les outils de sémantique du TAL, par exemple l'analyse distributionnelle, ont aussi montré de très bons résultats (Besançon *et al.*, 1999 ; Claveau et Kijak, 2015), avec l'avantage de pouvoir être appliqués sur les documents à indexer, et donc adaptés au domaine traité. Ces ressources sont souvent simplement adjointes au système de RI sans modification profonde de son fonctionnement. Ça peut être par extension de requête, ou, plus rarement, en modifiant la représentation du document (extension des documents, c'est-à-dire ajout des synonymes dès la phase d'indexation des documents, ou représentation par *synsets*), et parfois durant la phase de calcul de similarité (par exemple, par des techniques de *back-off* dans les modèles de langues). Bien connues en RI, les techniques de type Latent Semantic Indexing (LSI (Deerwester *et al.*, 1990)) ou Latent Dirichlet Allocation (LDA, (Hoffman *et al.*, 2010)) abordent également ce problème sémantique en proposant une représentation du document non plus dans un espace de mots, mais dans un espace de « concepts » (en fait, des combinaisons de mots apparaissant souvent dans les mêmes documents).

Pour aller plus loin que la sémantique lexicale, l'inclusion de connaissances riches et structurées sur le monde et le raisonnement à partir de ces connaissances sont également étudiés. Ce domaine, appelé RI sémantique, fait l'objet de l'article de Zargayouna *et al.* (2015) dans le numéro. La question de l'intégration de connaissances issues de bases structurées est aussi centrale dans les applications de questions-réponses présentées dans l'article de Grau *et al.* (2015).

4. Tâches mixtes

Les problématiques de l'accès à l'information contenue dans des documents textuels ne se résument pas à la seule tâche de recherche documentaire à l'aide d'une requête. Ainsi, d'autres applications sont aussi le point de rencontre entre TAL et RI. Nous en présentons quelques-unes ci-dessous, avant de nous attarder sur le cas emblématique de la RI translingue.

4.1. *Au-delà de la recherche documentaire*

La recherche de documents à partir d'une requête est l'application prototypique de la RI, mais l'accès à l'information peut prendre la forme d'autres applications. Il s'agit, par exemple, de l'extraction d'information (voir (Ananiadou *et al.*, 2013) pour un état de l'art), du résumé automatique mono- ou multidocument (Kundi *et al.*, n.d.),

3. C'est le cas notamment des travaux de Voorhees avec WordNet (Voorhees, 1994), qui, bien que démentis par de nombreuses autres études par la suite, restent largement cités comme preuve de l'inutilité des processus TAL en sémantique lexicale en RI.

ou encore des systèmes de questions-réponses, présentés dans l'article de Grau *et al.* (2015) dans ce numéro.

Les problèmes de représentation du contenu textuel, de calcul de similarité et la prise en compte des phénomènes linguistiques que nous avons évoqués se posent également dans ces applications. D'autres pistes sont plus spécifiques, comme par exemple la résolution d'anaphores, très utilisée dans ces applications (Vicedo et Ferrandez, 2000 ; Steinberger *et al.*, 2005) alors qu'elle ne l'est pas en recherche documentaire, ou encore la génération de texte, utile dans certaines tâches de résumé...

4.2. *RI translingue*

S'il y a un sous-domaine de RI qui nécessite les outils TAL, c'est bien la RI translingue (aussi appelée RI translinguistique). La RI translingue est née du besoin qu'un utilisateur peut vouloir retrouver des documents écrits en une autre langue que sa propre langue (langue de requête). Par exemple, l'utilisateur peut lire les documents dans une autre langue, mais ne connaît pas les mots appropriés pour formuler la requête ; ou bien les documents dans la langue source ne sont pas suffisants pour répondre à son besoin d'information, et il faut chercher dans une ou plusieurs autres langues. Par rapport à la RI traditionnelle dans la même langue (RI monolingue), la RI translingue ajoute la dimension de traduction : on doit traduire la requête (langue source) dans la langue des documents (langue cible), ou traduire les documents dans le sens inverse (Nie, 2010). Cette phase de traduction étant un des problèmes centraux en TAL, on retrouve en RI translingue une forme de synergie entre TAL et RI. Il s'agit bien d'une synergie et non pas d'une simple combinaison car, en plus de l'utilisation des systèmes de traduction automatique (TA) comme boîte noire pour produire une traduction, il y a aussi beaucoup de tentatives pour adapter la phase de traduction à la tâche de RI. Cette adaptation s'avère utile car la traduction d'une requête (ou d'un document) n'a pas pour but de produire un texte compréhensible par un être humain (comme dans le cas de TA en général), mais de produire des mots appropriés pour identifier des documents pertinents. Par exemple, la bonne séquence de mots, la bonne structure syntaxique et les accords qui sont importants en TA, n'ont pas beaucoup d'impact en RI translingue, étant donné que la traduction sera réduite en un sac de mots dépourvu de tous ces facteurs. L'aspect le plus important dans la traduction d'une requête (ou d'un document) est d'identifier de bons mots descripteurs. Cette observation a amené des chercheurs à utiliser seulement des modèles de traduction statistique (modèles IBM) (Kraaij *et al.*, 2003 ; Gao *et al.*, 2006) ou même un dictionnaire bilingue (Pirkola *et al.*, 2001 ; Levow *et al.*, 2005) dans la traduction d'une requête, en ignorant le modèle de langue, élément très utile en TA pour la production d'une phrase légitime en langue cible. Il faut noter qu'il s'agit bien d'une synergie : le modèle de traduction est souvent bien intégré dans un modèle de recherche (Kraaij *et al.*, 2003), et non pas utilisé dans une étape indépendante.

Un autre aspect important dans la traduction d'une requête est de pouvoir générer toutes les expressions possibles pour décrire le même sens. Contrairement à la TA

qui produit en général une seule phrase en sortie, la RI translingue a souvent besoin d'inclure plusieurs traductions possibles pour le même mot source. Ceci permet de produire naturellement un effet bénéfique d'extension de requête. Quand un modèle de traduction est intégré dans un modèle de recherche, on peut prendre en compte naturellement plusieurs traductions possibles pour chaque mot source, en tenant compte des probabilités de traduction. Quand on utilise un système de TA, il est aussi possible de produire les n meilleures traductions pour une requête, qui couvrent plus d'expressions en langue cible. La question est de savoir comment combiner ces multiples traductions dans le processus de recherche. Cet aspect a été l'objet des études de Ture *et al.* (2012) et Yanjun Ma et Wang (2012).

Un point très intéressant est que la RI translingue a aussi eu un impact sur la RI monolingue. Même quand la requête et les documents sont écrits dans une même langue naturelle, on considère souvent que le sous-langage des requêtes est bien différent du sous-langage des documents. En effet, les vocabulaires utilisés dans les requêtes sont souvent très différents de ceux des documents. Quand une requête est soumise, il est important de savoir quels autres mots dans le langage des documents peuvent décrire le même sens ou des sens similaires. Ce problème peut être formulé comme un problème de traduction dans la même langue. Le modèle de recherche, dit par traduction, a ainsi été proposé pour la RI monolingue (Berger et Lafferty, 1999). Ce modèle est maintenant largement répandu dans les moteurs de recherche : on peut considérer une requête et le titre d'un document sur lequel un utilisateur a cliqué comme une paire de textes parallèles. Les modèles de traduction entraînés sur ces données peuvent aider à améliorer la recherche (Gao *et al.*, 2010).

Ainsi, la RI translingue est une application qui oblige la RI et la TA à travailler ensemble, mais le recours aux modèles de traduction en RI monolingue ne relève pas de cette même obligation, mais de son intérêt pour gérer le problème du paraphrasage. On y voit ici un bel exemple dans lequel les techniques de TAL, parfois développées dans des contextes très différents, peuvent aider à améliorer la RI. N'est-ce pas ce genre de synergie qu'on aimerait voir plus souvent entre les deux domaines ?

5. Évolution et perspectives

5.1. *Regard sur le passé : un rapprochement lent*

En 2000, la revue TAL avait déjà consacré un numéro sur les liens entre RI et TAL (Jacquemin, 2000). Dans sa préface, Chr. Jacquemin revenait sur la citation de K. Spärck-Jones que nous rappelions dans l'introduction, tout en la modérant. Son constat était alors le suivant : l'apport du TAL en RI reste peu évident en général, mais il peut être bénéfique si les conditions suivantes sont remplies :

1) la tâche de RI doit nécessiter une représentation fine ;

2) les outils du TAL utilisés ne reposent pas sur « des représentations des connaissances riches dont l'adaptation en vraie grandeur est incertaine » ;

3) les outils du TAL n'induisent pas un coût calculatoire élevé.

Seize ans après, le constat général que nous dressons, et qui a motivé ce numéro, est toujours identique : la fertilisation croisée entre ces deux domaines reste assez pauvre, même si les articles de ce numéro montre le potentiel de l'interaction entre TAL et RI. Pour autant, notre lecture de la situation diffère de celle proposée alors par Chr. Jacquemin. Tout d'abord, il est frappant de noter que seul le sens *TAL pour la RI* était considéré, la RI étant vue uniquement comme une application, mais pas comme un ensemble de concepts et de techniques pouvant être utiles au TAL. Pourtant, comme nous l'avons rappelé en section 2, les apports des techniques de la RI au sein de processus de TAL sont bien réels, que ce soit, par exemple, avec les représentations vectorielles, les pondérations, ou encore avec les procédures d'évaluation. À ce titre, il nous semble important d'encourager la diffusion au sein de la communauté TAL des développements récents en RI pour continuer cette fertilisation. Ces développements concernent les points évoqués précédemment sur la représentation des documents, le calcul de similarité, et les méthodologies et données d'évaluation. On peut par exemple citer les nouvelles représentations des textes utilisées dans certains modèles de RI, mélangeant modèle de langue et réseaux d'inférences (Metzler et Croft, 2004 ; Strohman *et al.*, 2005), qui dépassent les limites des sacs de mots en offrant la possibilité de calculer des similarités en tenant compte de la proximité des mots, de phénomènes de synonymie, etc.

Les conditions 2 et 3 évoquées par Chr. Jacquemin portent sur la mise en œuvre en condition réelle, c'est-à-dire respectivement sur l'adaptabilité et la scalabilité. Concernant l'adaptabilité, la dépendance des outils du TAL à des systèmes experts ou des ressources créées manuellement a grandement diminué avec le développement des approches par apprentissage (supervisé, semi-supervisé ou non supervisé) et la mise à disposition de ressources généralistes ou spécialisées dans un grand nombre de langues. Cette évolution du TAL, que l'on schématise souvent par le passage des approches expertes ou symboliques aux approches statistiques, rend ce point moins prégnant qu'il a pu être à une certaine époque. Concernant le passage à l'échelle, là aussi l'évolution des ressources de calcul, reposant sur les capacités propres des machines (mémoire, CPU, GPU), et surtout sur le calcul distribué (grappe de calcul, *cloud*), a changé la donne. Il est maintenant possible, et même courant, d'exécuter des processus très lourds sur des grandes masses de textes en des temps compatibles avec les tâches de RI.

En revanche, la question de l'apport du TAL pour aider à représenter l'information (point 1) reste essentielle pour analyser les succès et les échecs du TAL pour la RI. Les outils et ressources de TAL ne peuvent bénéficier à la RI que s'ils apportent quelque chose vis-à-vis du problème de l'ambiguïté, ou de celui du paraphrasage. Un point crucial est que cette connaissance supplémentaire doit être intégrable dans le système de RI. Cela peut se faire parfois très simplement (par exemple par extension de requête, voir sections 3.2 et 3.3), mais nécessite parfois une connaissance fine des mécanismes de RI, voire une révision complète de ces mécanismes (modification de la représentation, du calcul de similarité, etc. ; voir section 3.2 et l'article de Zargayouna

et al. (2015)). Dans ce cas, les interactions fructueuses ne peuvent se faire qu'au travers d'une connaissance assez pointue des deux domaines ou d'échanges étroits entre les deux communautés.

5.2. *Regard vers le futur : des révolutions à venir ?*

Depuis peu, l'essor de nouvelles techniques d'apprentissage, notamment l'apprentissage profond (*deep learning*), dessine un nouveau paysage pour l'avenir du TAL et de la RI, et donc de leurs interactions. Ces nouvelles techniques attaquent en effet deux des points de convergence entre TAL et RI, c'est-à-dire la représentation du texte et le calcul de similarité.

Pour le premier point, il est bien sûr question d'apprentissage de représentations, qu'elles soient dites distribuées, continues ou spectrales, de plongements de mots (*word embeddings*), etc. Tous ces systèmes de représentation, réinventant l'analyse distributionnelle, permettent de dépasser beaucoup des contraintes de la représentation sac de mots : des mots proches sémantiquement seront proches dans l'espace de représentation, les régularités morphologiques (mise au pluriel, par exemple) se traduisent par des régularités géométriques, et certains raisonnements y sont possibles (sous la forme d'analogies, par exemple : `le veau est à la vache ce que le marcassin est à la laie`). Plusieurs problèmes sont encore ouverts, dont celui, important pour la RI, de savoir comment représenter le contenu d'un texte à partir de la représentation de chacun de ses mots, même si des propositions existent (Le et Mikolov, 2014). Un autre problème ouvert, spécifiquement discuté dans les articles de Grau *et al.* (2015) et Zargayouna *et al.* (2015) de ce numéro, est celui de la représentation conjointe des données textuelles et de données structurées issues, par exemple, de bases de connaissances.

Comme nous l'évoquions en section 2.2, un moteur de recherche, dont le cœur est de calculer des similarités entre textes, peut être vu comme un classifieur. Il est donc compréhensible que beaucoup de travaux cherchent désormais à apprendre ce classifieur. Le développement conjoint de méthodes d'apprentissage adaptées (*metric learning*, *learn to rank* ou désormais réseaux profonds...) mais surtout la mise à disposition de données (requêtes et documents associés, logs de moteurs de recherche sur le Web) et la disponibilité de puissance de calcul ont bien entendu été encore une fois des facteurs déterminants.

Enfin, il faut noter que cette mutation des pratiques ne concerne pas que le matériau textuel, mais d'une manière plus générale toutes les données non structurées, pour lesquelles les mêmes approches d'apprentissage sont employées. Cela interroge la spécificité du texte, de la langue, vis-à-vis d'autres types de données, et donc la spécificité des outils du TAL et de la RI vis-à-vis d'autres outils de fouille de données. Finalement, nous terminons cette préface avec cette prédiction, que le prochain numéro de TAL dédié à la RI pourra peut-être éclaircir : les frontières entre les corpus techniques du TAL et de la RI sont en train de disparaître, non pas parce que ces

communautés se rapprochent, mais parce qu'elles se fondent l'une et l'autre dans une communauté plus grande, celle des sciences des données.

6. Comité spécifique à ce numéro

Nous tenons à remercier Pascale Sébillot pour le suivi de ce numéro, les membres du comité permanent de la revue TAL, ainsi que les membres du comité spécifique à ce numéro :

– Jacques Savoy, université de Neuchatel, Suisse ;

– Abdelmajid Ben Hamadou, MIRACL, université de Sfax, Tunisie ;

– Jaap Kamps, université d'Amsterdam, Pays-Bas ;

– Fatiha Sadat, Département informatique, UQAM, Canada ;

– Lyne Da Sylva, École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada ;

– Brigitte Grau, LIMSI-CNRS, ENSIIE, France ;

– Lorraine Goeuriot, LIG, université de Grenoble Alpes, France ;

– Benjamin Piwowarsky, LIP6-CNRS, UPMC, France ;

– Catherine Berrut, LIG, université J. Fourier, Grenoble, France ;

– Olivier Ferret, CEA-LIST, France ;

– Mohand Boughanem, IRIT, université de Toulouse, France ;

– Xavier Tannier, université Paris-Sud, LIMSI-CNRS, France ;

– Éric Gaussier, LIG, université J. Fourier, Grenoble, France ;

– Haïfa Zargayouna, LIPN, université Paris 13, France ;

– Guillaume Cabanac, IRIT, université Toulouse 3, France ;

– Mathias Géry, Laboratoire Hubert Curien, université de Saint-Étienne, France ;

– Mathieu Roche, Cirad, UMR TETIS, Montpellier, France.

7. Bibliographie

Acosta O. C., Villavicencio A., Moreira V. P., « Identification and Treatment of Multiword Expressions Applied to Information Retrieval », *Proceedings of the Workshop on Multiword Expressions : From Parsing and Generation to the Real World*, MWE '11, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 101-109, 2011.

Ananiadou S., Friburger N., Rosset S. (eds), *Entités nommées*, vol. 54-2, 2013.

Aronson A. R., Lang F.-M., « An overview of MetaMap : historical perspective and recent advances », *JAMIA*, vol. 17, n° 3, p. 229-236, 2010.

Bellot P., Bogers T., Geva S., Hall M., Huurdeman H., Kamps J., Kazai G., Koolen M., Moriceau V., Mothe J., Preminger M., SanJuan E., Schenkel R., Skov M., Tannier X., Walsh D., *Information Access Evaluation. Multilinguality, Multimodality, and Interaction : 5th Inter-*

national Conference of the CLEF Initiative, CLEF 2014, Springer International Publishing, Sheffield, UK, chapter Overview of INEX 2014, p. 212-228, September, 2014.

Berger A., Lafferty J., « Information retrieval as statistical translation », *in* ACM (ed.), *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*, p. 222-229, 1999.

Besançon R., Rajman M., Chappelier J.-C., « Textual Similarities based on a Distributional Approach », *in Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA'99)*, p. 180-184, 1999.

Chapelle O., Chang Y., « Yahoo ! Learning to Rank Challenge Overview », *in* O. Chapelle, Y. Chang, T. Liu (eds), *Proceedings of the Yahoo ! Learning to Rank Challenge, held at ICML 2010, Haifa, Israel, June 25, 2010*, vol. 14 of *JMLR Proceedings*, JMLR.org, p. 1-24, 2011.

Claveau V., Kijak E., « Analyse morphologique non supervisée en domaine biomédical. Application à la recherche d'information », *Traitement Automatique des Langues*, vol. 54, n° 1, p. 13-45, October, 2013.

Claveau V., Kijak E., « Thésaurus distributionnels pour la recherche d'information et vice-versa », *Revue des Sciences et Technologies de l'Information - Série Document Numérique*, 2015.

Claveau V., Kijak E., Ferret O., « Improving distributional thesauri by exploring the graph of neighbors », *Proceedings of the International Conference on Computational Linguistics, COLING*, Dublin, Irlande, August, 2014.

Claveau V., Lefèvre S., « Topic segmentation of TV-streams by watershed transform and vectorization », *Computer Speech and Language*, vol. 29, n° 1, p. 63-80, 2015.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, 1990.

Gao J., He X., Nie J.-Y., « Clickthrough-Based Translation Models for Web Search : from Word Models to Phrase Models », *in* editor (ed.), *Proceedings of the CIKM conference*, p. 1139-1148, 2010.

Gao J., Nie J.-Y., Wu G., Cao G., « Dependence Language Model for Information Retrieval », *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, ACM, New York, NY, USA, p. 170-177, 2004.

Gao J., Nie J., Zhou M., « Statistical query translation models for cross-language information retrieval », , vol. 5, n° 4, p. 296-322, 2006.

Gaussier E., « Unsupervised Learning of Derivational Morphology from Inflectional Corpora », *Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, États-Unis, 1999.

Grau B., Ligozat A.-L., Gleize M., « Recherche d'information précise : vers des modèles hybrides exploitant des sources d'information structurées et non structurées », *Traitement Automatique des Langues*, vol. 56, n° 3, p. 75-99, 2015.

Grefenstette G., Semmar N., Elkateb-Gara F., « Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications », *Computational Approaches to Semitic Languages - Workshop Proceedings*, University of Michigan, p. 31-38, 2005.

Haddad H., Bechikh Ali C., « Performance of Turkish Information Retrieval : Evaluating the Impact of Linguistic Parameters and Compound Nouns », *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, CICLing 2014, Springer-Verlag New York, Inc., New York, NY, USA, p. 381-391, 2014.

Hearst M., « Text-tiling : segmenting text into multi-paragraph subtopic passages », *Computational Linguistics*, vol. 23, n° 1, p. 33-64, 1997.

Hoffman M., Bach F. R., Blei D. M., « Online Learning for Latent Dirichlet Allocation », *in* J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (eds), *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., p. 856-864, 2010.

Jacquemin C. (ed.), *Traitement automatique des langues pour la recherche d'information*, vol. 41-2, 2000.

Kraaij W., Nie J.-Y., Simard M., « Embedding Web-based statistical translation models in cross-language information retrieval », , vol. 29, n° 3, p. 381-419, 2003.

Kundi F. M., Asghar M. Z., Zahra S. R., Ahmad S., Khan A., A Review of Text Summarization, Technical Report n° 4, MAGNT Research Report (ISSN. 1444-8939), n.d.

Le Q. V., Mikolov T., « Distributed Representations of Sentences and Documents », *CoRR*, 2014.

Levow G.-A., Oard D. W., Resnik P., « Dictionary-based techniques for cross-language information retrieval », , vol. 41, p. 523-547, 2005.

Lovins J. B., « Development of a Stemming Algorithm », *Mechanical Translation and Computational Linguistics*, vol. 1, p. 22-31, 1968.

Maisonnasse L., Gaussier É., Chevallet J., « Modélisation de relations dans l'approche modèle de langue en recherche d'information », *Actes de la COnférence en Recherche d'Infomations et Applications - CORIA 2008*, Trégastel, France, March 12-14, 2008, p. 305-319, 2008.

Metzler D., Croft W., « Combining the Language Model and Inference Network Approaches to Retrieval », *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, vol. 40, n° 5, p. 735-750, 2004.

Moreau F., Claveau V., Sébillot P., « Automatic morphological query expansion using analogy-based machine learning », *Proceedings of the European Conference on Information Retrieval, ECIR'07*, Rome, Italie, avril, 2007.

Moreau F., Sébillot P., Contributions des techniques du traitement automatique des langues à la recherche d'information, Rapport de recherche n° 1690, IRISA, 2005.

Nie J.-Y., *Cross-Language Information Retrieval*, Synthesis Lectures on Human Language Technologies series, Morgan & Claypool, 2010.

Ollagnier A., Fournier S., Bellot P., « Recherche de livres par analyse de requêtes longues en langage naturel », *Traitement Automatique des Langues*, vol. 56, n° 3, p. 23-47, 2015.

Peng F., Huang X., Schuurmans D., Cercone N., « Investigating the Relationship Between Word Segmentation Performance and Retrieval Performance in Chinese IR », *Proceedings of the 19th International Conference on Computational Linguistics, COLING*, p. 1-7, 2002.

Pirkola A., Hedlund T., Keskustalo H., Järvelin K., « Dictionary-Based Cross-Language Information Retrieval : Problems, Methods, and Research Findings », , vol. 4, n° 3–4, p. 209-230, 2001.

Ponte J. M., Croft W. B., « A Language Modeling Approach to Information Retrieval », *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, Melbourne, Australia, p. 275-281, 1998.

Porter M., « An Algorithm for Suffix Stripping », *Program*, vol. 14, n° 3, p. 130-137, 1980.

Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *Proceedings of the 7th Text Retrieval Conference, TREC-7*, p. 199-210, 1998.

Salton G., McGill M., *Introduction to modern information retrieval*, McGraw-Hill, 1983.

Savoy J., « Stemming of French Words Based on Grammatical Categories », *Journal of the American Society for Information Science (JASIS)*, vol. 44, n° 1, p. 1-9, 1993.

Savoy J., Morphologie et Recherche d'Information, Rapport technique, Institut interfacultaire d'informatique, Université de Neuchâtel, 2002.

Shen W., Nie J.-Y., « Is Concept Mapping Useful for Biomedical Information Retrieval ? », *Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction : 6th International Conference of the CLEF Association, CLEF'15*, September, 2015.

Spärck-Jones K., « What is the Role of NLP in Text Retrieval ? », *in* T. Strzalkowski (ed.), *Natural Language Information Retrieval*, Kluwer Academic Publishers, p. 1-24, 1999.

Steinberger J., Kabadjov M. A., Poesio M., « Improving LSA-based summarization with anaphora resolution », *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 1-8, 2005.

Strohman T., Metzler D., Turtle H., Croft W., Indri : A language-model based search engine for complex queries (extended version), Technical report, CIIR, 2005.

Strzalkowski T., Lin F., Wang J., Perez-Carballo J., « Evaluating Natural Language Processing Techniques in Information Retrieval », *in* T. Strzalkowski (ed.), *Natural Language Information Retrieval*, Kluwer Academic Publishers, p. 113-145, 1999.

Ture F., Lin J., Oard D. W., « Combining Statistical Translation Techniques for Cross-Language Information Retrieval », *Proceedings of the International Conference on Computational Linguistics, COLING*, p. 2685-2702, 2012.

Vicedo J. L., Ferrandez A., « Importance of Pronominal Anaphora resolution in Question Answering systems », *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL*, p. 555-562, 2000.

Voorhees E. M., « Query Expansion Using Lexical-semantic Relations », *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, Springer-Verlag New York, Inc., New York, NY, USA, p. 61-69, 1994.

Yanjun Ma Jian-Yun Nie H. W., Wang H., « Opening Machine Translation Black Box for Cross-Language Information Retrieval », *Proceedings of the AIRS conference*, p. 467-476, 2012.

Zargayouna H., Roussey C., Chevallet J.-P., « Recherche d'information sémantique : état des lieux », *Traitement Automatique des Langues*, vol. 56, n° 3, p. 49-73, 2015.

Zhong Z., Ng H. T., « Word Sense Disambiguation Improves Information Retrieval », *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, ACL*, ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 273-282, 2012.