Recherche d'information et fouille de textes

Patrice Bellot, Brigitte Grau

Recherche d'information et fouille de textes

Patrice Bellot & Brigitte Grau

1-Introduction

Comprendre un texte est un but que l'Intelligence Artificielle (IA) s'est fixé depuis ses débuts et les premiers travaux apportant des réponses ont vu le jour dans les années 70s. Depuis lors, le thème est toujours d'actualité, bien que les buts et méthodes qu'il recouvre aient considérablement évolués. Il est donc nécessaire de regarder de plus près ce qui se cache derrière cette dénomination générale de « compréhension de texte ».

Les premiers travaux, qui ont eu lieu du milieu des années 70 jusqu'au milieu des années 80 [Charniak 1972; Dyer 1983; Schank et al. 1977], étudiaient des textes relatant de courtes histoires et comprendre signifiait mettre en évidence les tenants et aboutissants de l'histoire – les sujets traités, les événements décrits, les relations de causalité les reliant – ainsi que le rôle de chaque personnage, ses motivations et ses intentions. La compréhension était vue comme un processus d'inférence visant à expliciter tout l'implicite présent dans un texte en le retrouvant à partir des connaissances sémantiques et pragmatiques dont disposait la machine. Cela présupposait une modélisation préalable de ces connaissances. On rejoint ici les travaux effectués sur les différents formalismes de représentation des connaissances en IA, décrivant d'une part les sens associés aux mots de la langue (réseaux sémantiques *vs* logique, et notamment graphes conceptuels [Sowa 1984] et d'autre part les connaissances pragmatiques [Schank 1982].

Tous ces travaux ont montré leur limite dès lors qu'il s'agissait de modéliser manuellement ces connaissances pour tous les domaines, ou de les apprendre automatiquement. Le problème de la compréhension automatique en domaine ouvert restait donc entier.

Puisque le problème ainsi posé est insoluble en l'état des connaissances, une approche alternative consiste à le redéfinir et à le décomposer en sous-tâches potentiellement plus faciles à résoudre. Ainsi la compréhension de texte peut être redéfinie selon différents points de vue sur le texte qui permettent de répondre à des besoins spécifiques. De même qu'un lecteur ne lit pas un texte de façon identique selon qu'il veut évaluer sa pertinence par rapport à un thème qui l'intéresse (tâche de type recherche documentaire), qu'il veut classer des documents, prendre connaissances des événements relatés ou rechercher une information précise, de même les processus automatiques seront multiples et s'intéresseront à des aspects différents du texte en fonction de la tâche visée. Suivant le type de connaissance cherché dans un document, le lecteur n'extraira du texte que l'information qui l'intéresse et s'appuiera pour cela sur les indices et sur les connaissances qui lui permettent de réaliser sa tâche de lecture, et donc de compréhension, sans avoir à tout assimiler. On peut alors parler de compréhension à niveaux variables, qui va permettre d'accéder à des niveaux de sens différents.

Cette démarche est bien illustrée par les travaux en extraction d'information, évalués dans le cadre des conférences MUC [Grishman and Sundheim 1996], qui ont eu lieu de la fin des années 1980 jusqu'en 1998. L'extraction d'information consistait alors à modéliser un besoin d'information par un patron, décrit par un ensemble d'attributs typés, et à chercher à remplir ces attributs selon l'information contenue dans les textes. C'est ainsi que se sont notamment développées les recherches sur les « entités nommées » (à savoir le repérage de noms de personne, d'organisation, de lieu, de date, etc.) et sur les relations entre ces entités.

C'est aussi dans cette optique que se sont développées les approches se situant au niveau du document, que ce soit pour la recherche d'information ou pour en déterminer la structure

thématique, moins ambitieuses quant au type d'analyse, mais visant à traiter des textes portant sur n'importe quel domaine. Ces travaux exploitent des critères de surface, présents explicitement dans les textes. Ici, l'objet étudié est le texte, et non le processus de compréhension du texte.

L'auteur d'un texte dispose en effet d'un ensemble d'outils afin de le rendre compréhensible. Outre les mots qui rendent compte d'un contenu, l'auteur va construire un ensemble structuré et cohérent, en choisissant de développer ses idées dans des parties différentes qu'il organisera les unes par rapport aux autres. Afin d'expliciter la relation entre deux passages et deux idées, l'auteur utilise des *connecteurs*, des expressions linguistiques figées (ou semi-figées) et joue avec le temps des verbes. Pour mettre en évidence des points importants, il utilisera là encore des marques spécifiques. Toutes ces caractéristiques sont autant d'indices qui vont permettre une analyse des textes, fournissant un premier niveau de représentation du sujet traité.

Une propriété fondamentale d'un texte qui est largement exploitée pour la recherche de documents est sa cohésion lexicale. Le principe de cohésion lexicale est fondé sur le fait que le développement d'un thème entraîne la répétition, soit à l'identique soit par l'intermédiaire de variantes sémantiques, de mots qui lui sont caractéristiques, et cela dans une zone limitée du texte. Ainsi, en se fondant sur la récurrence, la distribution et les relations entre les mots d'un texte, on peut caractériser sa ou ses thématiques.

Enfin, à l'intersection de la recherche d'information et de l'extraction d'information, se situe la tâche de recherche de réponses à des questions. Dans le cas d'un besoin précis d'information, qui peut s'exprimer en une question (une question en langue naturelle portera plus d'informations qu'un ensemble de termes), on cherchera à fournir en résultat la réponse attendue, et non un document. Ainsi, si on veut savoir qui a tué Henri IV, il suffit de poser directement la question pour obtenir juste un nom, Ravaillac en l'occurrence, ce qui n'interdit pas d'y adjoindre la présence d'un document ou d'un passage justificatif. Dans ce type de tâche, la partie de texte à analyser est focalisée sur l'information à retrouver, et les processus mis en œuvre vont de méthodes exploitant uniquement des indices de surfaces à des méthodes de compréhension profonde des passages, exploitant des sources de connaissances sémantiques.

Toutes ces tâches peuvent être modélisées par des approches symboliques, des approches numériques ou des approches hybrides faisant collaborer les deux approches précédentes. Les approches symboliques relèvent d'une démarche analytique quant à l'usage de la langue pour la tâche étudiée, et visent la production de règles d'analyse. Les approches numériques exploitent la récurrence de phénomènes langagiers et modélisent leur combinaison. Les modèles sont capables de tenir compte de différents types d'informations, certaines sur plusieurs niveaux linguistiques. Force est de reconnaître que ces modèles numériques se comportent bien sur beaucoup de tâches avec en entrée des connaissances assez rudimentaires sur la langue. Les tâches très complexes requièrent une analyse linguistique plus élaborée, en amont des modèles lors de l'annotation des textes, qu'elle soit automatique ou manuelle, ou bien au cœur du modèle, et cela reste un problème ouvert.

Nous allons montrer, sur les différentes tâches qui nous intéressent dans cet article, quel niveau de sens est modélisé ou atteint par les différents types de processus, quels types de connaissances sont utilisés et quels sont les résultats obtenus. Aussi, nous allons en préambule introduire les différentes ressources linguistiques existantes, côté connaissances ou processus (section 2). Puis nous présenterons les méthodes développées pour la recherche d'information (section 3) et, enfin, nous aborderons l'extraction d'information (section 4) pour terminer en évoquant les étapes nécessaires pour extraire une réponse à des questions (section 5).

2-Ressources

La disponibilité de ressources lexicales, syntaxiques ou sémantiques ainsi que de collections de documents, annotés ou non, est un élément déterminant pour le traitement automatique des

langues, aussi bien pour être appréhendées en tant qu'objet d'étude que pour la construction de modèles informatiques, qu'ils soient symboliques ou numériques, et pour l'aide à l'annotation et l'analyse.

Comme nous le verrons plus loin, les premiers modèles de recherche d'information fonctionnaient à partir d'approches de surface s'apparentant à de la reconnaissance de formes et tout au plus avait-on besoin de disposer d'une collection de documents de référence sur laquelle il était possible d'estimer, pour un certain nombre de requêtes, des critères d'évaluation tels que la précision et le rappel [Cleverdon 1967; Harter and Hert 1997]. Ultérieurement, la mise au point des modèles probabilistes, et plus encore des moteurs de recherche du Web, ont nécessité de disposer de jugements de pertinence et de retours utilisateurs en grande quantité. Ce type particulier de ressources est difficilement accessible et seules les grandes entreprises du Web y ont généralement accès (AOL et Microsoft ont rendu public en 2006 et 2007 les *logs* associés à quelques dizaines de millions de requêtes mais ceux-ci demeurent difficiles d'accès et ne sont pas sans causer certains problèmes de confidentialité). Parallèlement à ces ressources qui permettent de paramétrer finement les modèles de recherche, éventuellement selon des profils utilisateurs, se pose la question de l'exploitation de ressources linguistiques et sémantiques. Quelles sont les ressources utiles et dans quel cadre le sont-elles ? De quelle manière peut-on les intégrer dans les modèles de recherche et, plus généralement, dans le processus de fouille de textes. Autant de questions qui n'ont toujours pas à l'heure actuelle de réponse claire sauf pour l'extraction d'information et la recherche d'information précise (questions-réponses) qui nécessitent une fouille en profondeur. Dans ces domaines, il est désormais difficile, pour ne pas dire impossible, de se passer des approches supervisées d'apprentissage automatique, lesquelles nécessitent bien sûr des corpus annotés (*cf.* sections 4 et 5).

Du côté des ressources linguistiques, bien sûr plus ou moins disponibles selon les langues considérées, on distingue les bases lexicales (lexiques de formes fléchies[1], lexiques spécialisés en différents domaines, lexiques avec représentations phonémiques et orthographiques [2], lexiques de formes courantes, par exemple dans des publications pour enfants[3], lexiques de toponymes[4], lexiques multilingues de noms propres[5], lexiques de mots porteurs d'opinions[6]… mais aussi des fréquences de n-grammes de mots ou de lettres[7]) éventuellement hiérarchiques telles que Wordnet[8] pour l'anglais et WOLF pour le français[9]. Des ressources sémantiques sont également disponibles. FrameNet[10] propose ainsi plus d'un millier de structures sémantiques conceptuelles associées à près de 200 000 phrases en anglais étiquetées, tandis que VerbNet[11] organise, décrit et associe à FrameNet plus de 6 000 verbes anglais. Si ce type de ressources est

[1] Morphalou fournit plus de 500 000 formes appartenant à plus de 68 000 lemmes du français (http://www.cnrtl.fr/lexiques/morphalou/).

[2] Lexique3 « *fournit pour 135 000 mots du français: les représentations orthographiques et phonémiques, la syllabation, la catégorie grammaticale, le genre et le nombre, les fréquences, les lemmes associés, etc.* » (http://www.lexique.org)

[3] http://www2.mshs.univ-poitiers.fr/novlex/

[4] http://www.geonames.org

[5] http://www.cnrtl.fr/lexiques/prolex/

[6] SentiWordNet associe à chaque groupe de synonymes (*synset*) en anglais de Wordnet trois indices de subjectivité positive, négative et neutre (http://sentiwordnet.isti.cnr.it).

[7] Google Ngram, disponible au téléchargement, propose de visualiser l'évolution de l'usage de n-grammes de mots au fil des années et ceci pour un grand nombre de langues dont le français. Ces statistiques ont été collectées à partir de la numérisation des livres du programme Google Books (http://books.google.com/ ngrams/).

[8] http://wordnet.princeton.edu/

[9] https://gforge.inria.fr/projects/wolf/

[10] https://framenet.icsi.berkeley.edu/

[11] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

particulièrement utile pour les systèmes de questions-réponses, de détection d'implication textuelle et, plus généralement, pour toute tâche mettant en œuvre une analyse sémantique [Palmer et al. 2010], leur usage est plutôt limité en recherche d'information (*i.e.* recherche de documents) où leur apport est encore peu significatif en moyenne [Fautsch and Savoy 2009].

Toutes ces ressources ne sont pleinement utilisables que grâce à une certaine standardisation des normes d'écriture et une généralisation de plateformes logicielles ouvertes pour les manipuler. C'est ainsi que l'on peut exploiter des étiqueteurs en parties du discours (par ex. TreeTagger[12] ou LIA_Tagg[13]), des raciniseurs (par ex. Snowball[14]), des correcteurs orthographiques (par ex. Aspell[15]), des extracteurs d'entités nommées (par ex. Stanford NE[16] pour l'anglais ou ACABIT pour le français[17]), des analyseurs syntaxiques (Stanford Parser[18] pour l'anglais, le chinois, l'arabe…, Xerox XIP[19] ou Sygmart[20] pour le français)… jusqu'à des plateformes complètes de traitement automatique des langues (par ex. Apache UIMA[21] et GATE[22]).

3-Recherche d'information

Un modèle de recherche d'information doit accomplir plusieurs rôles dont le plus important est de fournir un cadre théorique pour la modélisation de la mesure de pertinence associant une requête utilisateur et un document. Un document peut correspondre à un texte plus ou moins long, plus ou moins structuré, plus ou moins respectueux des bonnes règles d'écriture et des usages. Suivant son niveau de complexité, le modèle de recherche peut prendre en compte différents types de descripteurs et d'indices textuels, qu'il s'agisse de marqueurs lexicaux, syntaxiques, sémantiques ou structurels (voire prosodiques dans le cas de la recherche de documents audio), qu'ils soient obtenus via une analyse en surface ou une analyse en profondeur (dépendances lointaines, références). Certains modèles vont tenter de caractériser un document dans les limites du corpus dont il est issu en tenant compte d'indices globaux tels que la longueur moyenne des documents, ou encore la distribution fréquentielle moyenne des mots du lexique dans un corpus plus large. Les modèles seront alors plus ou moins aptes à répondre à des besoins différents en information qui, selon le contexte et le type de question posée, peuvent aller de la recherche de documents à la recherche d'informations précises, de la recherche de facettes spécifiques (localisation, temporalité…) à la recherche de la diversité la plus grande *etc*. Pour un besoin en information et un corpus de documents donnés, toute la question est de déterminer quel modèle et quels paramètres seront les plus efficaces en s'aidant, éventuellement, de telle ou telle ressource linguistique (dictionnaires de formes fléchies, thesaurus...) et de telle ou telle base de connaissances au sens le plus large du terme (ontologies lexicalisées, encyclopédies en ligne...).

Le modèle booléen est le plus simple des modèles de recherche d'information. C'est aussi le premier qui s'est imposé dans le monde de la recherche d'informations. Il est basé sur la théorie des ensembles et l'algèbre de Boole. Le modèle booléen considère que les termes de l'index sont présents ou absents dans un document. L'impact de la formulation de la requête est ici particulièrement fort. Soit la requête correspond à une conjonction de mots-clés et la forte

[12] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[13] http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html
[14] http://snowball.tartarus.org
[15] http://aspell.net
[16] http://nlp.stanford.edu/software/CRF-NER.shtml
[17] http://www.bdaille.com
[18] http://nlp.stanford.edu/software/lex-parser.shtml
[19] http://open.xerox.com/Services/XIPParser
[20] http://www.sygtext.fr
[21] http://uima.apache.org
[22] http://gate.ac.uk

précision induite peut se traduire par un silence élevé, soit il s'agit d'une disjonction aboutissant à un meilleur rappel mais à une précision affaiblie qui nécessitera des filtrages ultérieurs. Dans sa forme classique, le modèle booléen ne permet pas d'ordonner les documents trouvés en fonction de la question. En effet, si deux documents partagent le même nombre de mots communs avec la requête, seule une pondération non binaire de ces mots permet de différencier (et donc d'ordonner) ces documents [Salton et al. 1983] grâce au calcul de la valeur d'une fonction de score. D'une manière générale, la pondération permet d'établir, aussi bien dans les documents que dans la requête, un ordre d'importance entre les mots sur lequel se base ce calcul. Un exemple de pondération est donné ci-dessous. L'étude des pondérations est bien sûr particulièrement importante [Greiif 1998; Greiif et al. 2002; Li et al. 2008; Robertson and Sparck-Jones 1976].

Le modèle vectoriel représente les documents du corpus et les requêtes par des vecteurs de mots clés. Ces mots clés sont eux-mêmes extraits des textes lors d'une phase d'indexation et peuvent correspondre à des mots isolés, des mots-composés, ou à des chaînes de mots plus ou moins longues et caractéristiques. Pour chaque document, un poids est attribué à chacune des entrées de l'index qu'il contient. Dans le modèle vectoriel, le vecteur requête est représenté dans le même espace que les vecteurs documents. Le vecteur requête peut alors être comparé à chacun des vecteurs documents. Cette comparaison correspond au calcul de la valeur d'une fonction de similarité ou de distance (par exemple la distance euclidienne[23]) entre les vecteurs documents et le vecteur requête. Ces différentes valeurs permettent d'ordonner, en fonction de la requête, les documents trouvés. Dans le modèle vectoriel, les mots sont généralement supposés comme étant mutuellement indépendants. Ceci est une forte simplification puisqu'il est clair que la présence ou l'absence d'un mot dans un texte dépend des autres mots qui constituent ce texte. Dans une optique de recherche documentaire, un mot important est un mot qui permet de fortement caractériser un document par rapport à l'ensemble des documents du corpus cible. Autrement dit, un mot est important si sa sélection prioritaire par rapport aux autres mots candidats permet d'améliorer les résultats de la recherche. Il est raisonnable de penser que plus un mot apparaît dans un document, plus le sens de ce mot influe sur la thématique du document. A l'inverse, un mot sera fortement caractéristique d'un document s'il ne s'agit pas d'un mot présent dans un trop grand nombre d'autres documents. La conjonction de ces deux hypothèses revient à dire que, pour un document donné, un mot important est un mot qui est fréquent dans ce document mais qui ne se retrouve que dans un faible nombre d'autres documents. Cette définition correspond à introduire les facteurs TF (*term frequency*) et IDF (*inverse document frequency*). Même si ces dernières restent souvent utilisées, des pondérations plus efficaces que les classiques TF.IDF ont été proposées [Robertson and Sparck-Jones 1976; Savoy 2003; Savoy et al. 1997]. Elles différencient les mots en fonction de l'écart constaté dans leur usage au sein d'un document donné par rapport à leur usage moyen dans une collection (distribution fréquentielle).

D'autres modèles de recherche exploitent des retours fournis par les utilisateurs eux-mêmes sur les documents. On parle alors de modèle supervisé, alors que le modèle vectoriel précédent était non supervisé. Certaines déclinaisons [Robertson 1977] rentrent dans ce cadre et permettent de représenter le processus de recherche comme un processus de décision probabiliste : le coût, pour l'utilisateur, associé à la récupération d'un document doit être minimisé. Un document ne devrait être proposé à l'utilisateur que si le coût associé à sa lecture (temps passé par rapport au gain d'information engendré) est faible. Autrement dit, la règle de décision équivaut à proposer un document seulement si le rapport entre la probabilité qu'il soit pertinent et celle qu'il ne le soit pas est supérieur à un seuil donné. On cherche alors à modéliser l'ensemble des documents

[23] Si x et y sont deux vecteurs tels que $\vec{x} = (x_1,..,x_i,....,x_n)$ et $\vec{y} = (..,y_i,...)$ alors la distance euclidienne qui les sépare vaut $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.

pertinents ainsi que celui des documents non pertinents. La solution à ce problème peut passer par un processus itératif durant lequel l'utilisateur sélectionne à chaque itération des documents qu'il juge pertinents et qui sont utilisés pour générer automatiquement une nouvelle requête. Du fait de la structure en réseau des pages (hyperliens), les moteurs de recherche du Web exploitent en outre des indices non linguistiques : nombre de pages qui pointent vers une page donnée, probabilité d'accéder à cette page par une « promenade aléatoire » [Langville and Meyer 2006]… Ces indices sont à la base du fameux *PageRank* introduit au sein du moteur de recherche Google à la fin des années 1990 et qui considère qu'une page Web est d'autant plus importante que les liens qui y conduisent sont nombreux et proviennent eux-mêmes de pages importantes [Brin & Page, 1998].

Une autre extension du modèle vectoriel est le modèle LSI (*Latent Semantic Indexing*) [Deerwester et al. 1990]. Le but de LSI est de corriger les défauts du modèle vectoriel liés à la non-prise en compte des variations linguistiques et principalement de la synonymie. Ce modèle utilise les techniques de l'analyse en composante principale sur l'espace lexical afin de le projeter dans un espace plus conceptuel. Pour ce faire, LSI exploite les corrélations (co-occurrences) entre les mots afin de regrouper ceux qui sont susceptibles de représenter un même concept, ou un concept proche, dans une même classe de mots (par exemple *livre, monographie, ouvrage* peuvent être associés éventuellement en compagnie de *roman, nouvelle, étude* etc.). On obtient ainsi une représentation conceptuelle des documents, ce qui limite l'impact de la variation dans l'utilisation des mots dans les documents.

A la fin des années 1990, Ponte and Croft [1998] introduisent l'usage des modèles de langue probabilistes en recherche d'information. Un tel modèle tente de capter les régularités d'une langue (succession probable de mots ou de bi-grammes ou de tri-grammes de mots) en observant les co-occurrences lexicales dans un corpus d'entraînement. Un modèle de langue estime alors la vraisemblance (la probabilité *a posteriori*) d'avoir une séquence de mots donnée dans une langue donnée (ou pour un thème donné dans une langue donnée). Son utilisation en recherche d'information consiste à considérer que tout document est représenté par un modèle de langue ayant permis de générer son contenu. La pertinence d'un document vis-à-vis d'une requête est traduite par la probabilité que la requête puisse être générée à partir du même modèle de langue que le document.

Notons que les modèles qui viennent d'être présentés ne sont pas capables, sans l'usage de prétraitements ou de post-traitements, de prendre en compte les liens sémantiques entre les unités lexicales (synonymie ou proximité sémantique, liens hiérarchiques), pas plus que l'expression de négations. Des mesures sémantiques existent pourtant pour estimer à quel point un concept est éloigné d'un autre au sein de ressources sémantiques structurées et autres ontologies. Mais la façon d'intégrer ces approches aux modèles de surface précédents demeure un problème ouvert notamment en ce qui concerne leur application dans un cadre de recherche non spécialisée dans un domaine métier [Li, Zheng, Yang, Bu, Ge, Zhu, Zhang and Huang 2008]. Pour une interrogation dans un domaine de spécialité pour lequel des ressources sémantiques sont disponibles (lexiques, thesaurus, ontologies…), une recherche d'information dite sémantique procède souvent en deux phases : la première est destinée conceptualiser les documents et les requêtes afin de projeter l'espace lexical sur un espace conceptuel prédéfini (par exemple grâce au logiciel Metamap [Aronson and Lang 2010] et à une ontologie telle que SNOMED [Spackman et al. 1997] pour le domaine médical) et la seconde à estimer les scores de pertinence requête-documents en fonction des proximités sémantiques, dans la hiérarchie conceptuelle, des concepts liés aux documents et à la requête. Ces derniers scores peuvent être combinés avec ceux obtenus sans conceptualisation par application des méthodes de recherche vues plus haut. L'efficacité de la recherche d'information sémantique au sein de corpus de documents faiblement structurés et faiblement annotés dépend de la spécificité des requêtes et des ressources sémantiques utilisées. Les résultats des évaluations *Medical* et *e-Health* des

conférences TREC (*Text REtrieval Conference*) et CLEF (*Cross-Language Evaluation Forum*) montre qu'il s'agit bien là d'un problème toujours ouvert [Hamdan et al. 2013; Koopman et al. 2012; Limsopatham et al. 2012].

4) Extraction d'information

L'extraction d'information consiste à reconnaître certains types d'information dans des textes, afin de structurer l'information qu'ils contiennent en rajoutant des annotations ou en remplissant des bases de données ou de connaissances. L'extraction d'information couvre différentes tâches comme la reconnaissance d'entités nommées, la tâche la plus étudiée, l'extraction de relations ou plus récemment l'extraction d'événements. Une entité nommée (EN) est une unité lexicale que l'on cherche à reconnaître, typer et éventuellement normaliser, et qui fait référence à des objets particuliers du monde: personne, organisation, lieu, date, etc., ou des entités caractéristiques d'un domaine, comme des produits, des maladies, des traitements, etc. En cela, la frontière est floue entre entités nommées d'un domaine et termes d'un domaine. L'extraction de relations et d'événements consiste à reconnaître des relations entre entités, permettant ainsi de structurer l'information.

L'extraction d'information a fait l'objet d'évaluations depuis de nombreuses années, initiées par la création de la conférence MUC (*Message Understanding Conference*) en 1987. La tâche consistait à remplir des schémas prédéfinis portant sur un domaine particulier, par exemple le terrorisme, et par la suite différentes campagnes se sont spécialisées dans l'extraction d'EN avec ACE (*Automatic Content Extraction* [Doddington et al. 2004]), ou l'extraction de relations avec KBP (*Knowledge Base Population*) par exemple qui consiste à extraire de l'information à propos d'entités d'une base de connaissance, en domaine ouvert ou en domaine de spécialité. En domaine biomédical, le challenge LLL05 (*Learning Language in Logic*) [Nédellec 2005] portait sur l'extraction des interactions entre des protéines et des gènes et DDI Extraction [Segura-Bedmar et al. 2010] sur l'extraction des interactions entre médicaments. En 2010, le challenge i2b2 [Uzuner et al. 2011] proposait une tâche d'extraction de différentes relations entre des médicaments, des examens et des problèmes médicaux.

Les types d'entités nommées à reconnaître ont évolué au cours du temps, allant des cinq types définis dans MUC en 1995 (personne, organisation, lieu, date et expressions numériques) à un plus grand nombre dans Sekine and Nobata [2004], entités pouvant être hiérarchisées [Rosset et al. 2012].

La reconnaissance d'entités nommées (REN) s'appuie sur des critères de surface (majuscules, nombres *etc*.), des critères syntaxiques (notamment de la syntaxe locale), des listes de déclencheurs et des ressources sémantiques, notamment des listes de dénomination). Alors qu'aux début de la tâche de REN, on trouvait beaucoup d'approches performantes consistant à modéliser manuellement des règles, dorénavant, avec l'accroissement des corpus disponibles, les approches relèvent quasiment toutes d'un mécanisme d'apprentissage statistique : HMM [Zhou and Su 2002], SVM [Isozaki and Kazawa 2002], CRF [McCallum and Li 2003]. De ce fait, la connaissance linguistique est modélisée en partie explicitement dans les annotations *via* les guides d'annotation [Galibert et al. 2010], et dans les différents traits donnés aux modèles (par exemple catégorie morpho-syntaxique, structure syntaxique, classe sémantique, etc.). Ces derniers sont souvent d'assez bas niveau tout en permettant d'obtenir des résultats très élevés.

Les tâches les plus récentes concernent l'agrégation d'informations issues de différents textes, la normalisation d'entités sous-tendue par la reconnaissance d'anaphores, la reconnaissance d'entités hiérarchisées *etc*.

5) Répondre à des questions posées en langue naturelle

La recherche de réponses précises à des questions dans des textes pose des problèmes différents de ceux de la recherche de documents répondant à une requête. En effet, l'objectif est de fouiller automatiquement des documents à la recherche d'une réponse précise, ou d'un

élément de réponse, pour éviter ce travail à l'utilisateur. Ainsi, les systèmes de question-réponse (SQR) procèdent généralement en trois étapes : analyser (1) la question de manière à identifier une caractérisation de la réponse cherchée ainsi qu'une caractérisation des passages susceptibles de contenir cette réponse, sélectionner et analyser (2) ces passages afin d'en extraire (3) une réponse possible. Ainsi deux problèmes se posent : a) retrouver des passages qui contiennent la même information que celle donnée dans la question, quelle que soit la manière d'en parler, de l'énoncer ; et b) être capable de sélectionner l'élément qui répond à la question dans ces passages pertinents.

Trouver la réponse à des questions est aussi un moyen d'évaluer la compréhension d'un texte. C'est à cette fin que sont construits les tests de compréhension formulés sous forme de QCM (Questions à Choix Multiples) posé sur un texte et qui consiste à choisir une réponse parmi plusieurs. Les méthodes sont très proches de celles proposées dans les SQRs, à cela près que les candidats réponses sont connus.

Relations entre question et réponse

Un passage de texte pertinent pour en extraire une réponse peut être défini comme un ensemble de phrases, de l'ordre de une à trois phrases, qui contient l'information donnée dans la question et la réponse attendue. Généralement, cette information n'est pas exprimée dans les mêmes termes que ceux de la question et différents types de variations linguistiques sont à traiter entre la question et les extraits de textes. Au niveau lexical, les variations reposent sur l'emploi de :

- synonymes ou de relations sémantiques telles que l'hyperonymie et l'hyponymie pour désigner des entités ou des relations ;
- variations morphologiques, telles que la transformation verbe – nom
- combinaisons de ces variations

L'exemple de la figure 1 montre des exemples de telles variations.

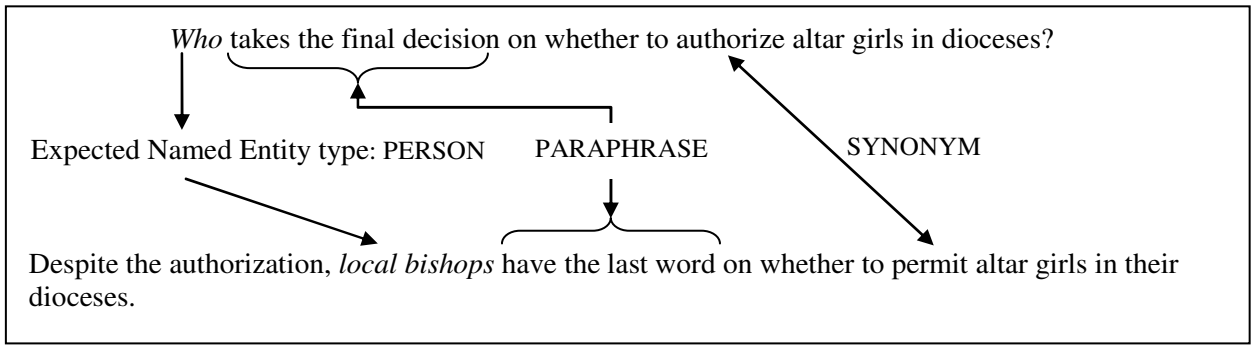

Figure 1: Variations lexicales entre la question et l'extrait de texte

Au niveau des extraits de texte, les systèmes doivent tenir compte de phénomènes d'anaphores et de paraphrases, de phrases ou de sous-phrases. La figure 2 illustre l'emploi d'anaphores, avec « their » qui reprend « Orville and Wilbur Wright ». Même si on retrouve l'information donnée dans la question presque à l'identique dans l'extrait, la relation de fraternité entre Orville et Wilbur n'est pas explicite (car l'expression « Orville and Wilbur Wright » ne suffit pas pour indiquer ce lien de parenté), et devrait être vérifiée, ou inférée, à partir d'autres documents, ou d'une base de connaissances, pour être certain qu'il s'agit des personnes cherchées.

Un extrait de texte sera donc considéré comme pertinent s'il contient une paraphrase de la question mise sous forme déclarative plus la réponse exacte. Généralement, les passages pertinents ne sont pas de strictes paraphrases ; ils contiennent l'information cherchée plus

d'autres, et de ce fait correspondent plus à une relation d'inclusion qu'à une stricte équivalence de sens, ou bien la réponse peut être déduite de l'information qu'ils décrivent, et l'on parle alors d'implication textuelle.

Le problème d'extraction d'une réponse peut alors être énoncé de manière générale comme le fait d'établir qu'un extrait de texte implique la réponse, représentée comme une hypothèse construite à partir de la reformulation déclarative de la question dans laquelle la place de la réponse attendue est marquée, et en instanciant cette place par la réponse candidate que l'on cherche à valider.

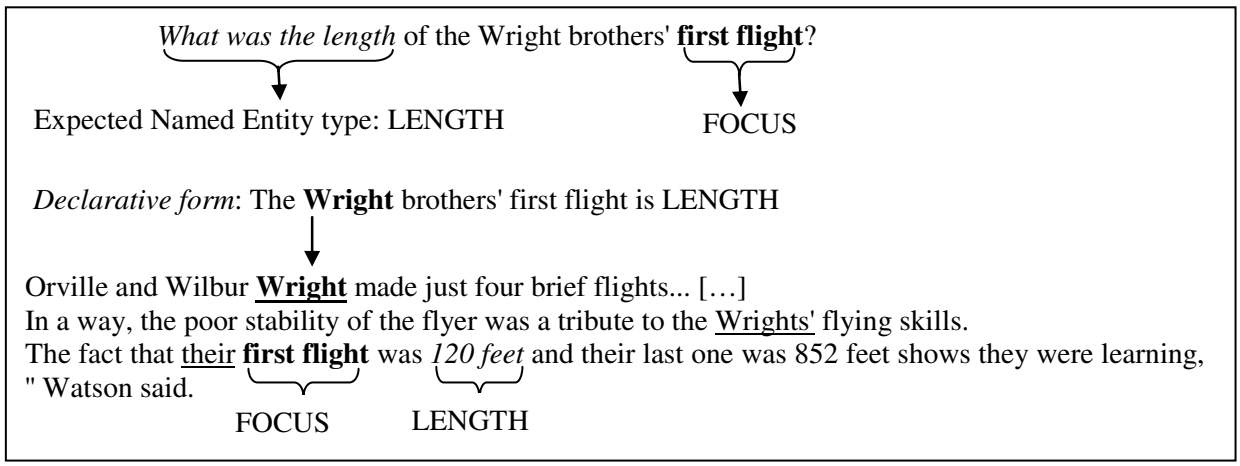

Figure 2: Phénomènes syntaxiques et discursifs entre la question et l'extrait de texte

L'implication textuelle

Les méthodes choisies vont être très liées à la base de connaissances dans laquelle s'effectue la recherche, texte ou base structurée ou semi-structurée. On peut différencier la recherche de réponses en domaine de spécialité *vs* la recherche de réponses en domaine ouvert, c'est-à-dire sur tous les sujets possibles, ainsi que la recherche sur le Web *vs* la recherche dans des collections de texte prédéfinies. Nous nous concentrerons ici sur les méthodes appliquées en domaine ouvert.

Les méthodes proposées reposent souvent sur l'apprentissage d'un classifieur dont le but est de décider s'il y a ou non implication textuelle à partir de différents critères permettant de décrire localement, ou séparément, des phénomènes ; leur combinaison, qui est difficile à décrire explicitement car elle relève souvent de processus de raisonnement complexes et variés, est assuré par le modèle de classification choisi (on retrouve souvent l'emploi dun modèle SVM[24] à cette fin). Les critères calculés sur les passages reprennent les variations linguistiques énoncées précédemment : termes communs, sous-phrases similaires, importance des termes communs, type sémantique de réponse attendu *vs* type du candidat réponse, redondance du candidat réponse dans des passages différents... [Grappy et al. 2011; Lin 2007; Magnini et al. 2002]. Afin de rapprocher termes et expressions, les systèmes font usage de différentes bases de connaissances sémantiques : WordNet [Miller et al. 1990] ou VerbOcean[25] [Chklovski and Pantel 2004], ou des bases de paraphrase [Clarke et al. 2003]. La reconnaissance de paraphrases de sous-phrase ou de l'hypothèse complète peut reposer sur des méthodes fondées sur des critères de surface (a), avec par exemple le calcul de la plus longue chaine commune [Newman et al. 2005], méthodes qui peuvent aussi intégrer des variations linguistiques [Bensley and Hickl 2008; Herrera et al. 2006; Ligozat et al. 2007], des méthodes reposant sur des correspondances

[24] Support Vector Machine

25 http://demo.patrickpantel.com/demos/verbocean/

entre structures syntaxiques (b) [Iftene and Moruz 2009; Kouylekov et al. 2006] ou entre représentations sémantiques (c) [Wang et al. 2009].

Néanmoins, des approches symboliques ont été explorées, et Moldovan et al. [2003] ont proposé dans le cadre des SQR un mécanisme de preuve logique, où les règles d'inférence ont été modélisées à partir des gloses[26] de WordNet pour constituer extendedWordNet [Mihalcea and Moldovan 2001]. Un modèle par preuve étant moins robuste que les approches par similarité, il a été utilisé conjointement avec ces dernières [Bensley and Hickl 2008; Clark and Harrison 2009; Tatu et al. 2006].

Dans le cadre des challenges RTE[27] permettant l'évaluation de l'implication textuelle, les meilleurs systèmes obtiennent environ 65-70 % de bonnes décisions. En QR, les meilleures performances sont de l'ordre de 70 %, voire 90 %. Le système WATSON développé par IBM [Ferrucci et al. 2010] et qui a joué au jeu américain Jeopardy[28] a réussi à gagner contre des spécialistes de ce jeu. Il met en œuvre énormément de méthodes différentes, s'appuie sur de nombreuses bases de connaissances et un très grand nombre de documents.

De manière générale, les bonnes performances des SQR reposent sur l'utilisation de nombreuses ressources et de la mise en œuvre de différentes méthodes. Il est difficile d'en retirer une modélisation explicite des raisonnements élaborés, mais ces approches permettent de faire collaborer différents niveaux de connaissances sur la langue, calculés de manière approchée ou précise sur des phrases ou portions de phrase.

7- Conclusion

Les différentes approches de recherche et d'extraction d'information que nous avons présentées soulignent la complexité des tâches qui nous sont devenues familières depuis l'usage généralisé des moteurs de recherche de l'Internet. Si les progrès de ces dernières années sont notables pour la fouille du Web, de nombreux points demeurent très perfectibles car les performances des systèmes récents sur des collections documentaires fermées sont comparables à ceux que l'on obtenait il y a une quinzaine d'années [Armstrong et al. 2009]. On citera pour exemple la grande faiblesse des systèmes de recherche actuels à prendre en compte les négations, les modes et les temps des verbes, à résoudre les références, à évaluer le sens en contexte, à relier des informations dispersées dans plusieurs phrases, paragraphes ou documents, à estimer la crédibilité d'une information… Face à toutes ces difficultés, les propositions théoriques sont multiples et tentent de combiner efficacement — du point de vue des performances qualitatives et quantitatives — approches symboliques et numériques, apprentissage automatique et statistique, reconnaissance des formes et logique, ressources et bases de connaissances. Les expérimentations à grande échelle sont nombreuses, sans cesse plus complexes et sur des données plus diverses et volumineuses : les tâches de recherche d'information liées à des entités nommées au sein des *Text REtrieval Conferences* (TREC) utilise des corpus de plus de 20 tera-octets, soit 20 000 fois plus volumineuses que les corpus utilisés il y a dix ans ! Les modèles de recherche se doivent désormais de fonctionner sur des flux de textes et non plus sur des collections uniquement statiques, être suffisamment robustes pour considérer à la fois des textes relus et corrigés, des micro-messages sur des réseaux sociaux et des transcriptions automatiques de parole. La personnalisation et la contextualisation de la recherche d'information sont deux autres aspects cruciaux : puisqu'il n'est pas possible de *tout* indiquer dans une requête, les modèles doivent exploiter d'autres indices que ceux fournis explicitement par l'utilisateur au moment de sa recherche. Etre capable de formuler des hypothèses sur ce que recherche un utilisateur et de fouiller finement des masses de documents (en entendant par document toute production linguistique disponible à un instant donné) sans

[26] Une glose fournit la définition d'un concept, ou un exemple d'usage
[27] Recognizing Textual Entailment
[28] Jeopardy, jeu télévisé aux Etats-Unis, où les candidats doivent trouver la question, à partir de réponses

cesse croissante et mouvante est un enjeu majeur pour les années à venir. La croisée des disciplines et des approches, qu'elles soient issues de l'informatique, de la linguistique, des mathématiques, des neuro-sciences et des sciences cognitives, sont autant de pistes prometteuses. Mais il se peut aussi que des méthodes *a priori* simplistes accomplissent des prouesses. Il ne reste qu'à les découvrir.

Bibliographie

ARMSTRONG, T.G., MOFFAT, A., WEBBER, W. AND ZOBEL, J. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management* ACM, 601-610.

ARONSON, A.R. AND LANG, F.-M. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association 17*, 229-236.

BENSLEY, J. AND HICKL, A. 2008. Workshop: Application of LCC's GROUNDHOG System for RTE-4.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In *7th WWW Conference*, Brisbane, Australia.

CHARNIAK, E. 1972. Toward a model of children's story comprehension Cambridge, MIT.

CHKLOVSKI, T. AND PANTEL, P. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP*, 33-40.

CLARK, P. AND HARRISON, P. 2009. An inference-based approach to recognizing entailment. *Proc. of TAC*.

CLARKE, C.L.A., CORMACK, G.V., KEMKES, G., LASZLO, M., LYNAM, T.R., TERRA EGIDIO, L. AND TILKER, P.L. 2003. Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002). In *The Eleventh Text Retrieval Conference (TREC 2002)*.

CLEVERDON, C. 1967. The Cranfield tests on index language devices. In *Aslib proceedings* MCB UP Ltd, 173-194.

DEERWESTER, S.C., DUMAIS, S., LANDAUER, T.K., FURNAS, G.W. AND HARSHMAN, R.A. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science 41*, 391-407.

DODDINGTON, G., MITCHELL, A., PRZYBOCKI, M., RAMSHAW, L., STRASSEL, S. AND WEISCHEDEL, R. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of LREC* Citeseer, 837-840.

DYER, M.G. 1983. *In-depth understanding: A computer model of integrated processing for narrative comprehension*. MIT press.

FAUTSCH, C. AND SAVOY, J. 2009. Evaluation de diverses stratégies de désambiguïsation lexicale. In *Actes 6ème Conférence en Recherche d'Information et Applications CORIA'09*.

FERRUCCI, D., BROWN, E., CHU-CARROLL, J., FAN, J., GONDEK, D., KALYANPUR, A.A., LALLY, A., MURDOCK, J.W., NYBERG, E. AND PRAGER, J. 2010. Building Watson: An overview of the DeepQA project. *AI magazine 31*, 59-79.

GALIBERT, O., QUINTARD, L., ROSSET, S., ZWEIGENBAUM, P., NÉDELLEC, C., AUBIN, S., GILLARD, L., RAYSZ, J.-P., POIS, D. AND TANNIER, X. 2010. Named and specific entity detection in varied data: The Quaero named entity baseline evaluation. *Proc. of LREC, Valletta, Malta. ELRA*.

GRAPPY, A., GRAU, B., FALCO, M.-H., LIGOZAT, A.-L., ROBBA, I. AND VILNAT, A. 2011. Selecting answers to questions from web documents by a robust validation process. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01* IEEE Computer Society, 55-62.

GREIIF, W.R. 1998. A theory of term weighting based on exploratory data analysis. In *Proceedings of the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia1998 ACM Press, 11-20.

GREIIF, W.R., MORGAN, W.T. AND PONTE, J.M. 2002. The role of variance in term weighting for probabilistic information retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management* ACM Press, McLean, Virginia, USA, 252-260.

GRISHMAN, R. AND SUNDHEIM, B. 1996. Message understanding conference-6: A brief history. In *Proceedings of COLING*, 466-471.

HAMDAN, H., ALBITAR, S., BELLOT, P., ESPINASSE, B. AND FOURNIER, S. 2013. LSIS at TREC

2012 Medical Track–Experiments with conceptualization, a DFR model and a semantic measure. In *Proceedings of the The Twenty-First Text REtrieval Conference (TREC 2012)*, Gaithersburg (USA)2013 NIST.

HARTER, S.P. AND HERT, C.A. 1997. Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. *Annual Review of Information Science and Technology (ARIST) 32*, 3-94.

HERRERA, J., RODRIGO, A., PENAS, A. AND VERDEJO, F. 2006. UNED submission to AVE 2006. In *Workshop CLEF 2006*.

IFTENE, A. AND MORUZ, M.-A. 2009. Uaic participation at rte5. *Proceedings of TAC, Gaithersburg, Maryland*.

ISOZAKI, H. AND KAZAWA, H. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-* Association for Computational Linguistics, 1-7.

KOOPMAN, B., ZUCCON, G., BRUZA, P., SITBON, L. AND LAWLEY, M. 2012. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management* ACM, 2439-2442.

KOUYLEKOV, M., NEGRI, M., MAGNINI, B. AND COPPOLA, B. 2006. Towards Entailment-based question-answering. In *Working notes of the CLEF 2005 workshop*.

LANGVILLE, A.N. AND MEYER, C.D. 2006. *Google-s PageRank and Beyond*. Princeton University Press.

LI, F., ZHENG, Z., YANG, T., BU, F., GE, R., ZHU, X., ZHANG, X. AND HUANG, M. 2008. Thu quanta at tac 2008 qa and rte track. In *Proceedings of Human Language Technologies Conference/Conference on Empirical methods in Natural Language Processing (HLT/EMNLP), Vancouver, BC, Canada*.

LIGOZAT, A.-L., GRAU, B., VILNAT, A., ROBBA, I. AND GRAPPY, A. 2007. Towards an automatic validation of answers in Question Answering. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on* IEEE, 444-447.

LIMSOPATHAM, N., MCCREADIE, R., ALBAKOUR, M.-D., MACDONALD, C., SANTOS, R.L. AND OUNIS, I. 2012. University of Glasgow at TREC 2012: Experiments with Terrier in Medical Records, Microblog, and Web Tracks TREC.

LIN, J. 2007. An exploration of the principles underlying redundancy-based factoid question answering. *ACM Trans. Inf. Syst. 25*, 4-53.

MAGNINI, B., NEGRI, M., PREVETE, R. AND TANEV, H. 2002. Is it the right answer?: exploiting web redundancy for Answer Validation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* Association for Computational Linguistics, 425-432.

MCCALLUM, A. AND LI, W. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* Association for Computational Linguistics, 188-191.

MIHALCEA, R. AND MOLDOVAN, D.I. 2001. extended wordnet: Progress report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources* Citeseer.

MILLER, G.A., BECKWITH, R., FELLBAUM, C., GROSS, D. AND MILLER, K. 1990. Introduction to WordNet : An Online Lexical Database. *International Journal of Lexicography 3*, 235-244.

MOLDOVAN, D., CLARK, C., HARABAGIU, S. AND MAIORANO, S. 2003. Cogex: A logic prover for question answering. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* Association for Computational Linguistics, 87-93.

NÉDELLEC, C. 2005. Learning language in logic-genic interaction extraction challenge. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)* Citeseer.

NEWMAN, E., STOKES, N., DUNNION, J. AND CARTHY, J. 2005. UCD IIRG approach to the textual entailment challenge. In *the PASCAL Recognising Textual Entailment Challenge Workshop* Citeseer, 53-56.

PALMER, M., GILDEA, D. AND XUE, N. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies 3*, 1-103.

PONTE, J.M. AND CROFT, W.B. 1998. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* ACM Press, Melbourne, Australia, 275-281.

ROBERTSON, S.E. 1977. The probability ranking principle in IR. *Journal of documentation 33*, 294-304.
ROBERTSON, S.E. AND SPARCK-JONES, K. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 129-146.
ROSSET, S., GROUIN, C., GALIBERT, O., ZWEIGENBAUM, P., FORT, K. AND QUINTARD, L. 2012. Les entités nommées dans le programme QUAERO. In *Programme QUAERO*.
SALTON, G., FOX, E. AND WU, H. 1983. Extended boolean information retrieval. *Communications of the ACM 31*, 1002-1036.
SAVOY, J. 2003. Modèles en recherche d'information. In *Assistance intelligente à la recherche d'informations*, E. GAUSSIER AND M.-H. STÉFANINI Eds. Hermès, Paris, 31-70.
SAVOY, J., LE CALVÉ, A. AND VRAJITORU, D. 1997. Report on the TREC-S Experiment: Data Fusion and Collection Fusion. *NIST SPECIAL PUBLICATION SP*, 489-502.
SCHANK, R., ABELSON, R. AND SCHANK, R.C. 1977. *Scripts Plans Goals*. Lea.
SCHANK, R.C. 1982. Reminding and Memory Organization: An Introduction to MOPs. In *Strategies for natural language processing*, L.E.R. (ED.) Ed. Lawrence Erlbaum.
SEGURA-BEDMAR, I., MARTÍNEZ, P. AND DE PABLO-SÁNCHEZ, C. 2010. Extracting drug-drug interactions from biomedical texts. *BMC Bioinformatics 11*, P9.
SEKINE, S. AND NOBATA, C. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of LREC*.
SOWA, J.F. 1984. *Conceptual structures: information processing in mind and machine*. Addison Wesley.
SPACKMAN, K.A., CAMPBELL, K.E. AND CÔTÉ, R.A. 1997. SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium* American Medical Informatics Association, 640-644.
TATU, M., ILES, B., SLAVICK, J., NOVISCHI, A. AND MOLDOVAN, D. 2006. Cogex at the second recognizing textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 104-109.
UZUNER, Ö., SOUTH, B.R., SHEN, S. AND DUVALL, S.L. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association 18*, 552-556.
WANG, R., ZHANG, Y. AND NEUMANN, G. 2009. A joint syntactic-semantic representation for recognizing textual relatedness. *Proc. of TAC*.
ZHOU, G. AND SU, J. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* Association for Computational Linguistics, 473-480.