Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés

Adil El Ghali[1, 2] Daniel Hromada[1] Kaoutar El Ghali
(1) LUTIN UserLab, 30, avenue Corentin Cariou, 75930 Paris cedex 19
(2) IBM CAS France, 9 rue de Verdun, 94253 Gentilly
elghali@lutin-userlab.fr

RÉSUMÉ
Cet article présent le système hybride et multi-modulaire d'extraction des mots-clés à partir de corpus des articles scientifiques. Il s'agit d'un système multi-modulaire car intègre en soi les traitements 1) morphosyntaxiques (lemmatization et chunking) 2) sémantiques (Reflective Random Indexing) ainsi que 3) pragmatiques (modélisés par les règles de production). On parle aussi d'un système hybride car il était utilisé -sans modification majeure- pour trouver des solutions aux toutes les deux pistes du DEFT 2012. Pour la Piste 1 - où une terminologie était fournie - nous obtînmes le F-score de 0.9488 ; pour la Piste 2 – où aucune liste des mots clés candidates n'était pas fourni au préalable – le F-score obtenu est 0.5874.

ABSTRACT
Enriching and reasoning on semantic spaces for keyword extraction

This article presents a multi-modular hybrid system for extraction of keywords from corpus of scientific articles. System is multi-modular because it integrates components executing transformations on 1) morphosyntactic level (lemmatization and chunking) 2) semantic level (Reflected Random Indexing), as well as upon more 3) « pragmatic » aspects of processed documents, modeled by production rules. The system is hybrid because it was able to address both tracks of DEFT2012 competition – a «reduced search-space» scenario of Track 1, whose objective was to map the content of a scientific article upon one among the members of a « terminological list » ; as well as more « real-life » scenario of Track2 within which no list was associated to documents contained in the corpus. In both Tracks, the system hereby presented has obtained the an F-score of 0.9488 for the Track1, and 0.5874 for the Track2.

MOTS-CLÉS : Extraction de mots-clés, Espaces sémantiques, RRI, Réseau bayésien, Règles de production, Chunking.

KEYWORDS: Keyword extraction, Semantic spaces, RRI, Bayesian Network, Production Rules, Chunking.

1 Introduction

L'édition 2012 du défi fouille de textes (DEFT) a pour thème l'identification automatique des mots-clés indexant le contenu d'articles publiés dans des revues scientifiques. Deux pistes ont été proposées : dans la première (Piste 1) la terminologie des mots-clés est fournie, alors que dans la deuxième (Piste 2) l'attribution des mots-clés devait se faire sans terminologie.

Pour la réalisation de cette tâche nous avons décidé, dans la continuité de ce que nous avions réalisé en 2011 (El Ghali, 2011), de représenter le sens des termes et des documents du corpus dans des espaces sémantiques utilisant la variante *Reflective Random Indexing* (RRI). Le choix de RRI une variante de *Random Indexing* (RI) (Sahlgren, 2006) est motivé par les bonnes propriétés de cette méthode, héritées de RI et qui sont largement décrites dans la littérature (Cohen *et al.*, 2010a). Mais une de ces propriétés moins connue et commentée s'est révélée particulièrement pertinente pour le problème posé dans le cadre de cette édition du DEFT, à savoir l'uniformité de l'espace sémantique : en effet, les vecteurs construits par RRI pour représenter les documents et les termes du corpus sont « comparables ».

Dans la méthode que nous avons développé pour cette édition du DEFT, nous avons voulu répondre à deux questions principales :

1. quel serait l'apport d'un pré-traitement linguistique de surface aux espaces sémantiques ? et en quoi pourrait-on comparer ces pré-traitements aux méthodes de constructions d'espaces sémantiques permettant de capturer des éléments de structure ?
2. peut-on améliorer les méthodes de *scoring* développées dans les précédentes éditions du DEFT en utilisant les dernières avancées en Intelligence artificielle, notamment le raisonnement à base de règles et les graphes probabilistes, encodant respectivement des règles générales sur le choix des mots-clés et des informations incertaines issues du corpus d'apprentissage ?

La première question s'imposait naturellement du fait qu'une grande partie des mots-clés qui ont été fournis pour la Piste 1 sont en fait des groupes de mots et que leurs catégories morphosyntaxiques et grammaticales respectait des règles assez simples. Pour pouvoir traiter les mots-clés composés de plusieurs mots, certaines méthodes de représentation de textes en espaces sémantiques telles que BEAGLE (Jones et Mewhort, 2007), PSI (Cohen *et al.*, 2009), ou encore RRI avec des indexes positionnels (Widdows et Cohen, 2010), permettent d'encoder les informations sur l'ordre des mots. La deuxième question est née du fait que l'on disposait d'informations de nature différentes qui pouvait aider à attribuer correctement des mots-clés : sur la sémantique, sur la distribution des mots-clés, sur la structure, sur les revues dont sont issues les articles ... Ces informations pouvaient être difficilement encodées dans un seul formalisme de décision. Nous avons donc décidé de définir une procédure de décision pour l'attribution de mots-clés qui combine des règles symboliques avec des réseaux bayésiens, avec les *Règles de production Probabilistes* (Aït-Kaci et Bonnard, 2011).

Nous avons fait le choix d'aborder les deux pistes du défi de cette année de manière sensiblement identique, les mêmes méthodes ont été utilisées pour les deux pistes. Pour ce faire, nous avons construit une terminologie pour la Piste 2. Cette terminologie est une liste de mots-clés candidats établie en utilisant un espace sémantique et un pré-traitement linguistique de surface.

L'article est organisé comme suit : nous commençons par présenter dans la section 2 une analyse du corpus et des informations qui peuvent en être extraite et qui sont utiles pour la tâche d'attribution de mots-clés. Ensuite, dans la section 3, nous rappelons brièvement le principe de fonctionnement de RRI, puis nous décrivons comment incorporer les informations issue du pré-traitement linguistique dans les espaces sémantiques, mais aussi comment la liste des candidats mots-clés pour la Piste 2 est construite. Dans la section 4 nous présentons le principe de fonctionnement de la procédure de décision pour l'attribution des mots-clés. Enfin, dans la section 5 nous détaillons les caractéristiques de chacune des exécutions et discutons les résultats avant de conclure.

2 Le Corpus

2.1 Statistiques générales de corpus d'apprentissage

2.1.1 Piste 1

Pour la Piste 1, il y a 140 documents dans le corpus d'apprentissage. Les documents proviennent de 4 revues différentes, l'identificateur de la revue étant encodé dans le nom du fichier XML contenant l'article.

La liste terminologique – i.e. la liste contenant tous les termes uniques choisies comme un mot clé pour un document dans le corpus - associée au corpus d'apprentissage contient $T_{appr} = 666$ termes uniques.

Les nombres des mots-clés associés sont fournis pour chaque document du corpus d'apprentissage aussi bien que du corpus de test. En somme, $\Sigma_i N_{appr_i} = 754$. En moyenne, chaque article de corpus d'apprentissage a :

$\mathsf{mean}(N_{appr}) = 5.386$; $\mathsf{median}(N_{appr}) = 5; \mathsf{min}(N_{appr}) = 1; \mathsf{max}(N_{appr}) = 13; \mathsf{sd}(N_{appr}) = 1.344$

Etant donné que $\Sigma_i N_{appr_i} > T_{appr}$, il est évident qu'il y a des termes qui sont définis comme mots clés pour plusieurs articles. Le principe de bijection 1 terme – 1 article n'est pas donc applicable. Plus précisément, pour le corpus d'apprentissage, 604 mots clés sont associés à un seul article, 46 en sont associés à deux, 10 à trois, quatre mots clés (i.e. « identité », « interprétation », « enseignement de la traduction », « traduction ») sont chacun associés à quatre articles, tandis que le terme « humanitaire » est défini comme mot clé pour cinq articles et le terme « mondialisation » pour sept articles.

On note aussi que parmi 62 termes qui sont associés à plus qu'un article, seulement 26 (i.e. 41,9%) sont associés aux articles appartenants à plus qu'une revue.

Les analyses fréquentielles préliminaires montrent aussi que dans 141 parmi 740 cas, le mot clé ne se trouve pas dans le corps ni résumé d'article auquel il est associé. En d'autres termes, pour plus que 19% des mots clés, la fréquence de leur occurrence dans l'article est zéro, c'est donc plus qu'évident qu'il faut aller au-delà des fréquences « brutes » si on veut que notre système d'extraction des mots clés ait la précision > 80% (la Figure 1 montre les fréquences d'occurrence des mots-clés dans les documents associés).

L'objectif de la Piste 1 est donc de concevoir le système qui, partant de fichiers de corpus d'apprentissage contenant $D_{appr} * T_{appr} = 140 * 666 = 93240$ couplages (document, terme) serait capable à déterminer les couples ayant été établis par les auteurs de leurs documents.

2.1.2 Piste 2

Le corpus d'apprentissage contient 142 documents. Contrairement à la Piste 1, aucune liste terminologique n'est fournie, l'espace de recherche dans lequel on cherche les candidats censé d'être les mots clés est donc beaucoup plus grande. Mais les quantité des mots clés associés au différents articles sont présents. Grâce à ces quantités fournis dans la balise `<nombre>` des documents XML, on sait sans regarder au fichier de référence que la distribution de $\Sigma_i N_{appr_i} = 763$

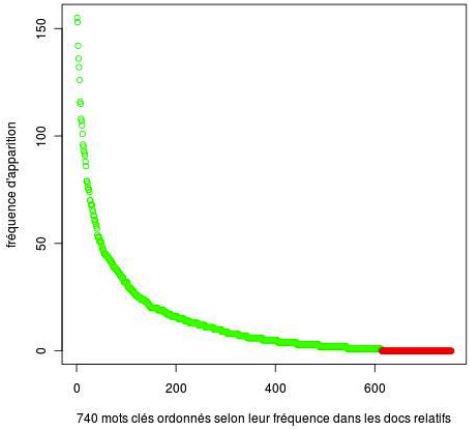

FIGURE 1 – Cca 19% (en rouge) des mots clés de corpus d'apprentissage ne figurent pas dans les documents auxquels ils sont attribués

associations entre mots clés et articles dispose de propriétés suivantes :

$\mathsf{mean}(N_{appr}) = 5.411; \mathsf{median}(N_{appr}) = 5; \mathsf{min}(N_{appr}) = 3; \mathsf{max}(N_{appr}) = 13; \mathsf{sd}(N_{appr}) = 1.404.$

L'analyse de fichier de référence révèle que parmi 681 termes qui couvrent l'ensemble de tous les mots clés du corpus d'apprentissage de piste2 , 627 en sont associés à un seul article, 37 à deux, 12 à trois, deux termes à (« humanitaire » et « didactique ») à quatre articles, les termes « identité » et « culture » étant associé à cinq articles et le terme « traduction » à huit documents. Étant donné que l'information concernant l'appartenance d'un article à une revue est présente, on sait aussi que parmi 54 termes associés à plus qu'un article, seulement 18 (i.e. 33.3%) sont associés à plus qu'une revue.

L'analyse des fréquences de mots clés dans les articles associés donne les résultats qui vont dans le même sens que ceux de la Piste 1 : dans 145 cas (19%), les mots clés n'apparaîssent pas dans l'article auquel ils étaient associés !

2.2 Statistiques générales du corpus de test

2.2.1 Piste 1

Le corpus de test de la Piste 1 contient $D_{test} = 94$ documents dans . La liste terminologique du corpus de test contient 478 termes uniques. Parmi ces 478 termes-candidats, 435 en sont associés avec un seul document, 34 aux deux documents différentes, quatre termes sont associés aux trois articles, et quatre termes aux quatre articles, le terme le plus réussi comme mot clé étant « identité » lui-même associé au six articles. Parmi les 43 termes associés à plus d'un article, 20 (i.e. 46,5%) sont associés aux articles appartenants à plus d'une revue.

La distribution de la somme du nombre des mots clés associés aux articles du corpus de test de la

Piste 1 ($\Sigma_i N_{test_i} = 537$) dispose de propriétés suivantes :

$\text{mean}(N_{test}) = 5.712; \text{median}(N_{test}) = 5; \text{min}(N_{test}) = 1; \text{max}(N_{test}) = 12; \text{sd}(N_{test}) = 1.701.$

2.2.2 Piste 2

La distribution de $\Sigma_i N_{test_i} = 484$ mots clés attribués aux 93 documents contenus dans le corpus de test de la Piste 2 est caractérisé par les mesures suivantes :

$\text{mean}(N_{test}) = 5.204; \text{median}(N_{test}) = 5; \text{min}(N_{test}) = 2; \text{max}(N_{test}) = 10; \text{sd}(N_{test}) = 1.323.$

La consultation des fichiers de référence obtenus après la fin de la phase competitive de DEFT2012 nous permets à savoir que parmi 35 termes associés à plus qu'un article, seulement 10 (i.e. 28,6%) sont associés aux articles appartenants à plus d'une revue.

2.3 Que peut-on apprendre d'autre du corpus ?

Un rapide parcours du corpus de d'apprentissage et de la terminologie fournie pour la Piste 1, nous montre qu'au delà des fréquences, les mots-clé choisis par les auteurs respectent quelques règles :

– les mots-clés sont différents entre eux : les auteurs n'utilisent que rarement des mots-clés très proches ;
– ils sont assez souvent repris dans l'introduction et la conclusion de l'article ;
– leur catégorie morphosyntaxique ou grammaticale est très rarement « verbale », les mot-clés les plus utilisés sont des noms (communs ou propres), des adjectifs ou des groupes nominaux ;

Par ailleurs, comme on pouvait s'y attendre les mots-clés sont fortement liés sémantiquement au document, comme le montre la figure 2 :

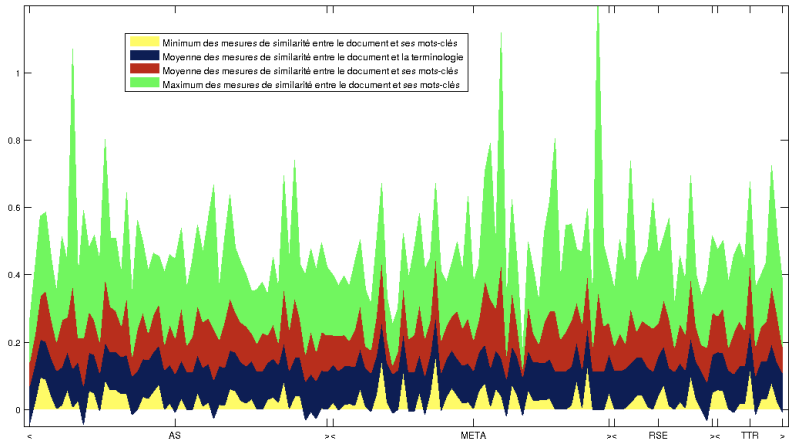

FIGURE 2 – Similarités document-mots-clés (min, max, mean) vs. document-terminologie (mean)

3 Espaces sémantiques

Les modèles de représentation vectorielle de la sémantique des mots sont une famille de modèles qui représentent la similarité sémantique entre les mots en fonction de l'environnement textuel dans lequel ces mots apparaissent. La distribution de co-occurrence des mots dans le corpus est rassemblée, analysée puis transformée en espace sémantique dans lequel les mots sont représentés comme des vecteurs dans un espace vectoriel de grande dimension. LSA (Landauer et Dumais, 1997), HAL (Lund et Burgess, 1996) et RI (Kanerva *et al.*, 2000) en sont quelques exemples. Ces modèles sont basés sur l'hypothèse distributionnelle de (Harris, 1968) qui affirme que les mots qui apparaissent dans des contextes similaires ont un sens similaire. La caractérisation de l'unité de contexte est une problèmatique commune à toutes ces méthodes, sa définition est différente suivant les modèles. Par exemple, LSA construit une matrice mot-document dans laquelle chaque cellule a_{ij} contient la fréquence d'un mot i dans une unité de contexte j. HAL définit une fenêtre flottante de n mots qui parcourt chaque mot du corpus, puis construit une matrice mot-mot dans laquelle chaque cellule a_{ij} contient la fréquence à laquelle un mot i co-occure avec un mot j dans la fenêtre précédemment définie.

Différentes méthodes mathématiques permettant d'extraire la signification des concepts, en réduisant la dimensionnalité de l'espace de co-occurence, sont appliquées à la distribution des fréquences stockées dans la matrice mot-document ou mot-mot. Le premier objectif de ces traitements mathématiques est d'extraire les «patrons» qui rendent compte des variations de fréquences et qui permettent d'éliminer ce qui peut être considéré comme du « bruit ». LSA emploie une méthode générale de décomposition linéaire d'une matrice en composantes indépendantes : la décomposition de valeur singulière (SVD). Dans HAL la dimension de l'espace est réduite en maintenant un nombre restreint de composantes principales de la matrice de co-occurrence. À la fin de ce processus de réduction de dimensionnalité, la similitude entre deux mots peut être calculée selon différentes méthodes. Classiquement, la valeur du cosinus de l'angle entre deux vecteurs correspondant à deux mots ou à deux groupes de mots est calculée afin d'approximer leur similarité sémantique.

3.1 *Reflective Random Indexing*

La méthode de construction d'espace sémantique utilisée est Reflective Random Indexing (RRI) (Cohen *et al.*, 2010a), c'est une nouvelle méthode de construction d'espaces sémantiques basée sur la projection aléatoire qui est assez différente des autres méthodes de construction d'espaces sémantiques. Ses particularités sont (i) qu'elle ne construit pas de matrice de co-occurrence et (ii) qu'elle ne nécessite pas, contrairement aux autres modèles vectoriels de représentation sémantique, des traitements statistiques lourds comme la SVD pour LSA. RRI est basée sur la projection aléatoire (Vempala, 2004; Bingham et Mannila, 2001), qui permet un meilleur passage à l'échelle pour grand nombre des documents. La construction d'un espace sémantique avec RRI se déroule comme suit :

– Créer une matrice $A(d \times n)$, contenant des vecteurs indexes, où d est le nombre de documents ou de contextes et n le nombre de dimensions choisies par l'expérimentateur. Les vecteurs indexes sont des vecteurs creux générés aléatoirement.
– Créer une matrice $B(t \times n)$, contenant des vecteurs termes, où t est le nombre de termes différents dans le corpus. Initialiser tous ces vecteurs avec des valeurs nulles pour démarrer la

construction de l'espace sémantique.
– Pour tout document du corpus, chaque fois qu'un terme τ apparaît dans un document δ, accumuler le vecteur index de δ au vecteur terme de τ.

à la fin du processus, les vecteurs termes qui apparaîssent dans des contextes similaires ont accumulé des vecteurs indexes similaire.

L'aspect « *Reflective* » dans RRI consiste à rejouer plusieurs cycles des trois étapes de l'algorithme non plus à partir de vecteurs aléatoires mais à partir des vecteurs indexes obtenues pour les documents. Ces cycles permettent de gommer l'aspect aléatoire de l'espace, le système convergeant généralement au bout d'un nombre réduit de cycles.

3.1.1 Semantic Vectors

Plusieurs implémentations libre de RRI sont disponibles, nous utilisons la librairie Semantic Vectors [1] (Widdows et Cohen, 2010). Semantic Vectors présente un certain nombre d'avantages par rapport aux autres librairies implémentant RRI, en particulier, parce qu'il offre, d'une part, une implémentation de RRI basé sur des indexes positionnels (Cohen *et al.*, 2010a) qui construit l'espace sémantique non plus en se basant sur les occurrences d'un terme dans un document mais dans une fenêtre glissante à la manière de HAL, cette version de RRI permet de capturer outre les informations sur la sémantiques des termes, des informations structurelles sur leur proximité. D'autre part, Semantic Vectors implante un certain nombre de mesures de similarité entre des groupes de mots, en particulier (i) la « disjonction quantique » (Cohen *et al.*, 2010b) qui permet de construire un volume correspondant à plusieurs termes dans l'espace sémantique et de calculer la distance entre ce volume et d'autres termes ou documents de l'espace ; (ii) « similarité tensorielle » qui prend en entrée une suite ordonnée de termes et calcule sa similarité avec d'autres suites ordonnées, exploitant ainsi les informations d'ordre provenant des indexes positionnels.

Semantic Vectors est utilisé dans nombre d'applications. Nous l'avons utilisé dans nos participations au DEFT depuis l'édition 2009. Dans des tâches proches de celle qui nous occupe, la librairie a été utilisée pour comparer RRI à d'autres méthodes d'espaces sémantiques pour la recherche de relations entre termes dans un corpus (Rangan, 2011).

3.2 Enrichir les espaces sémantiques avec des informations linguistiques

Dans le problème d'attributions de mots-clés à un texte, les termes utilisés comme mots-clés sont, pour une partie d'entre-eux, des groupes de mots. La sémantique associée à un groupe de mots dans espace sémantiques n'est pas aussi précise que celle associé à un mot : elle comprend des composantes de ce mots dans d'autres contextes. Pour pouvoir traiter la sémantique de ces groupes de mots, certaines méthodes de représentation du sens en espaces sémantiques telles que BEAGLE (Jones et Mewhort, 2007), PSI (Cohen *et al.*, 2009), ou encore RRI avec des indexes positionnels (Cohen *et al.*, 2010b; Widdows et Cohen, 2010), permettent d'encoder les informations sur l'ordre des mots. Nous avons voulu tester une autre méthode basée sur une analyse linguistique de surface du texte.

1. `http://code.google.com/p/semanticvectors/`

Le principe de cette méthode est d'identifier des groupes de mots candidats dans le texte via une phase de *chunking* (Abney, 1991) puis de construire des classes d'équivalence de chunks qui regroupent une majorité de mots identiques (après lemmatisation des mots) et qui sont sémantiquement proches - en se basant sur la sémantique, dans un espace sémantique "classique", des mots qu'ils contiennent -. Le corpus est alors transformé en remplaçant tous les chunks d'une même classe d'équivalence par un représentant de la classe et un nouvel espace sémantique est construit à partir de ce nouveau corpus, dans cet espace les représentants des classes de chunks sont considérés comme des mots.

Pour les besoins de la Piste 1, le *chunker* a été entrainé pour considérer comme chunk tous les mots-clés composés de la terminologie fournie. Dans la Piste 2 ce même chunker, ainsi que la procédure de construction de classes de chunks, sont utilisés pour construire une liste de mots-clés candidats.

4 Affectation de mots-clés comme procédure de décision mixte

4.1 Réseau Bayésien pour l'affectation de mots-clés

En analysant un corpus d'articles, nous cherchons, dans un premier temps, à déterminer la taille des différents mots-clés rattachés à un article donné. Dans un second temps, nous nous efforçons d'établir les probabilités d'appartenance de ces mots-clés à une liste pré-établie. Nous disposons pour chaque document du corpus des informations suivantes :

- les longueurs du résumé l et du texte L ;
- la revue R dans laquelle l'article est paru ;
- le nombre de mots-clés n et leurs tailles respectives $n_1, \ldots, n_n$ (*ie* le nombre de mots les composant) ;
- les similarités avec la totalité du lexique des mots-clés $(d_1, \ldots, d_N)$ (N taille de la terminologie) ;
- les mots-clés $(kw_1, \ldots, kw_n)$.

Il s'agit donc de trouver des relations entre les variables exogènes $(l, L, R, n, d_1, \ldots, d_N)$ permettant de prévoir le comportement des variables endogènes $(n_1, \ldots, n_n, kw_1, \ldots, kw_n)$. A cette fin, il faut disposer d'un formalisme de modélisation des connaissances adapté. Les réseaux bayésiens (Barber, 2012), étant des modèles graphiques auxquels sont associées des représentations probabilistes sous-jacentes, apparaissent comme particulièrement adaptés à notre cas d'étude.

Un réseau bayésien B est un couple (G, θ) où G est un graphe acyclique dirigé dont les noeuds représentent un ensemble de variables aléatoires $X = \{X_1, \ldots, X_n\}$ et $\theta_i = [P(X_i / C(X_i))]$ est la matrice des probabilités conditionnelles du nœud i connaissant l'état de ses parents $C(X_i)$.

L'intérêt des réseaux bayésiens est donc que leurs structures graphique et probabiliste permettent de prendre en charge une représentation modulaire des connaissances, une interprétation à la fois quantitative et qualitative des données. En effet, le graphe d'un réseau bayésien permet ainsi de représenter schématiquement les relations entre les variables du système à modéliser et les distributions de probabilités, elles, permettent de quantifier ces relations.

Le modèle que l'on se propose de construire est un réseau bayésien à variables discrètes (le nom de la revue R, les mots-clés kw_i, leur nombre n, leurs tailles n_i) et à variables continues (longueurs du résumé l, de l'article L et les similarité à la terminologie). C'est un modèle mixte, appelé modèle conditionnel gaussien, pour lequel la distribution des variables continues conditionnellement aux variables discrètes est une gaussienne multivariée. Cela implique qu'il peut y avoir des arcs partant de noeuds discrets vers des noeuds continus, mais pas l'inverse hormis pour le cas où les noeuds continus sont observables (ce qui est notre cas). Notons également que le nombre de variables $n_1, \ldots, n_n$ et $kw_1, \ldots, kw_n$ varie selon le nombre de mots-clés n ; le nombre de noeuds dans un réseau bayésien étant fixe, nous nous proposons de poser $n_1, \ldots, n_{25}$, les tailles des différents mots-clés avec $n_i = 0$ si $i > n$ et $kw_1, \ldots, kw_{25}$ les différents mots-clés avec $kw_i = NULL$ si $i > n$.

Pour résumer nous disposons des variables aléatoires suivantes représentées par les noeuds du réseau bayésien que l'on cherche à construire :

- R, le nom de la revue (variable discrète pouvant prendre 4 valeurs) ;
- l, la longueur du résumé (variable continue) ;
- L, la longueur de l'article (variable continue) ;
- n, le nombre de mots-clés (variable discrète pouvant prendre 25 valeurs) ;
- $n_1, \ldots, n_{25}$, la taille des mots-clés (variable discrète pouvant prendre 11 valeurs) ;
- $d_1, \ldots, d_{1062}$, les similarités à l'ensemble des mots-clés (variable continue) ;
- $kw_1, \ldots, kw_{25}$, les mots-clés (variable discrète pouvant prendre 1062 valeurs).

L'observation des distributions des documents entre les différentes revues nous permet d'affirmer que celles-ci sont similaires dans le corpus d'apprentissage et celui de test ; ce qui implique que le biais qu'introduit cette distribution n'impactera pas les performances du modèle à construire.

Les moyennes des longueurs de résumé l et d'article L présentent le même ordre de grandeur. Ces moyennes ne sont certes pas similaires dans le corpus d'apprentissage et celui de test, mais elles sont distribuées de la même manière, *ie* que les longueurs de résumé (respectivement d'article) sont égales dans le corpus d'apprentissage et dans celui de test au même facteur près. Notons également que les longueurs d'article et de résumé ne sont pas distribuées de la même manière ; cela veut dire qu'en plus de la relation directe évidente entre ces deux variables, il existe probablement une cause commune aux deux, ce qui se traduit dans la structure du réseau bayésien par la présence d'un parent commun.

Les distributions des nombres de mots par article (respectivement par résumé) peuvent être approximées par des mélanges de gaussiennes. Ces histogrammes sont similaires pour le corpus entier et pour celui d'apprentissage. Ce qui nous montre que l'échantillon étudié peut être considéré comme représentatif du problème. Toutefois, la relative disparité observée entre le corpus de test et celui d'apprentissage créera probablement un problème de biais qu'il faudra prendre en compte durant la construction du modèle.

Les histogrammes des nombres de mots par article (respectivement par résumé) représentent pour les différentes revues des distributions différentes. Ces variables sont donc directement reliées à la nature de la revue. Ces différentes distributions ont des formes quelconques, cependant, nous remarquons que l'on pourra les approximer par un mélange de gaussiennes ; ce qui nous conforte dans le choix d'un modèle conditionnel gaussien pour représenter ces variables dans un réseau bayésien.

En observant la monotonie des moyennes des similarités à la terminologie des mots-clés sur les différentes parties du corpus, nous remarquons qu'elle présente la même allure (et même quasiment le même tracé) dans tous les cas (corpus entier, corpus d'apprentissage, revue en particulier, ...). Cela nous permet de supposer que la sélection de mots-clés se fait strictement de la même manière partout, et donc l'idée d'en faire un modèle mathématique est parfaitement cohérente.

Sur la base de ces différentes observations, prenons un exemple de structure de réseau bayésien reliant les variables de notre problème. Par convention, les variables discrètes sont représentées par des noeuds carrés, les variables continues par des noeuds ronds et les variables observables par des noeuds ombrés (figure 3).

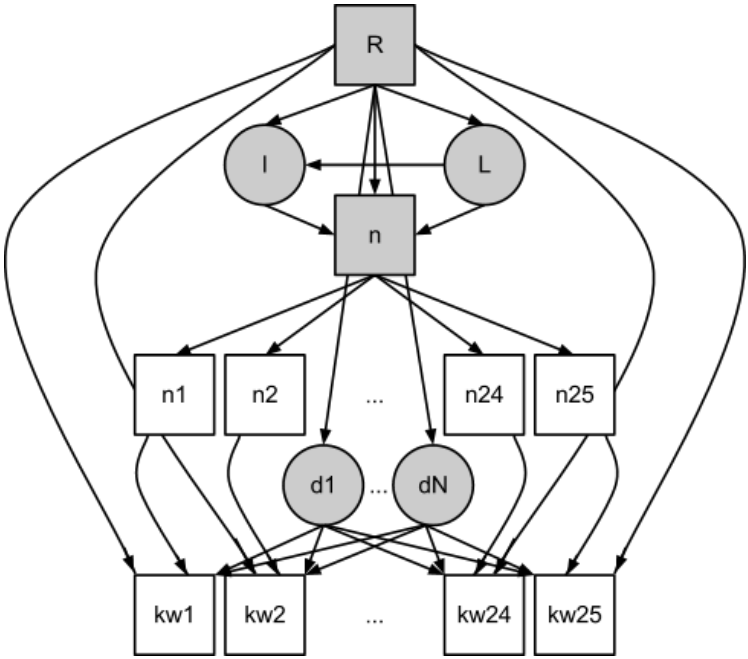

FIGURE 3 – Structure du réseau bayésien appris sur le corpus

4.2 Combiner des décisions statistiques avec du raisonnement à base de règles

Les récents travaux en intelligence artificielle sur la combinaison de méthodes de décision statistiques et de raisonnement à base de règles de production, comme les *Règles de Production Probabilistes* (PPR) de (Aït-Kaci et Bonnard, 2011), nous offrent un cadre pour modéliser une procédure de décision qui prend en compte ce qui est appris par le réseau bayésien décrit ci-dessus, et les connaissances symboliques encodées dans les règles sur le choix des mots-clés dont nous avons donné des exemples en 2.3.

Le principe de fonctionnement du système de décision, construit en se basant sur PPR, est de calculer un score pour chacun des mots-clés pour un document donné. Ce calcul est réalisé en utilisant des règles pouvant faire appel au réseau bayésien. Par exemple, la règle "*les mots-clés sont différents entre eux*" peut se traduire par la règle production "*si deux mots-clés sont proches alors augmenter le score de celui qui est le plus haute probabilité d'être un mot-clé du document et réduire l'autre*" qui s'écrit :

```
IF similarity(kw1, kw2) > seuil AND bnproba(kw1|doc) > bnproba(kw2|doc)
THEN increase-score(kw1, doc) AND decrease-score(kw2, doc)
```

Le système de règles que nous avons utilisé contient une quinzaine de règles. Nous ne pouvons pas les détailler ici par manque de place.

5 Les exécutions soumises

La table 1 résume les exécutions soumises par notre équipe. Ses résultats sont très satisfaisants pour toutes les approches que nous avons utilisé. La moyenne de F-score pour la Piste 1 pour l'ensemble des participants étant de 0,3575 et pour la Piste 2 de 0,2045. On notera que les premières exécutions pour les deux pistes (1.1 et 2.1) qui sont nos exécutions de base donnent des résultats corrects en des temps relativement bas.

Run	Precision	Rappel	F-score	Temps (en s)
1.1	0.4618	0.4618	0.4618	2
1.2	0.9479	0.9497	**0.9483**	7590
1.3	0.7486	0.7486	0.7486	-
2.1	0.2438	0.2438	0.2438	26
2.2	0.3471	0.3471	0.3471	269
2.3	0.5879	0.5867	**0.5873**	12700

TABLE 1 – Résultats soumis : performance et temps d'éxecution

5.1 Piste 1

5.1.1 Run 1.1 – *baseline* : RRI et k-NN

Dans cette exécution qui constitue notre *baseline*, nous avons construit un espace sémantique RRI avec l'ensemble des documents du corpus (appr + test), un document étant constitué par la concaténation du résumé et du corps de l'article. Puis pour chaque document d du corpus de test, nous avons retenu comme mots-clés les k plus proches voisins du document dans la terminologie, k étant le nombre de mots-clés pour le document d. Le vecteur pour un mot-clé kw_i composé des mots $w_1, ..., w_n$ étant obtenu en sommant les vecteurs des mots qu'il contient.

$$k\vec{w}_i = \Sigma_i \vec{w}_i \tag{1}$$

5.1.2 Run 1.2 – RRI(chunks), BN et règles

Dans cette exécution, qui a obtenu le meilleur résultat, nous avons construit un espace sémantique "enrichi" comme nous l'avons décrit dans la section 3.2, mais dans lequel un document était représenté par quatre vecteurs, un pour le résumé, un pour le corps de l'article et deux vecteurs pour le premier et le dernier paragraphe de l'article (que nous avons pris comme approximation de l'introduction et la conclusion) . Nous avons ensuite appris le réseau bayésien décrit en 4.1 en utilisant les distances entres les documents et les mots-clés obtenues sur cet espace. Enfin, nous avons utilisé la procédure de décision décrite en 4.2 pour affecter un score à chacun des mots-clés, les mots-clés retenus sont les k ayant les plus hauts scores (k étant le nombre de mots-clés pour le document).

5.1.3 Run 1.3

Dans le cadre de ce run, on a combiné les résultats de run 1 et run 2, en donnant une légère préférence aux candidates-termes lesquels sont plus longues que d'autres termes-candidates. On a donc combiné, par exemple, les termes-candidates de run1 :

> *Catalogne ; Narotzky ; conflit ; contexte ;district industriel ; femmes ; production traductionnelle ; production écrite ; réseau*

avec les termes-candidates de run 2 :

> *Espagne ; Narotzky ; anthropologie économique ; district industriel ; féminisme ; histoire ; réseaux de production ; économie politique ; économie régionale*

pour obtenir la liste des candidates de run3 :

> *district industriel ; réseaux de production ; économie politique ; production traductionnelle ; anthropologie économique ; Narotzky ; économie régionale ; production écrite ; féminisme*

Le score du candidat était calculé par la formule :

$$score = F_r * (l - F_a) \tag{2}$$

où F_r est la fréquence relative du terme-candidat dans l'article analysé, F_a est la fréquence absolue du terme-candidat dans tous les articles du corpus et l est le nombre de caractères du terme-candidat.

5.2 Piste 2

5.2.1 Run 2.1 – *baseline* : RRI et k-NN

Cette exécution est identique à la première exécution de la Piste 1 5.1.1, la terminologie obtenue par la méthode décrite en 3.2 contient 3000 candidats mots-clés.

5.2.2 Run 2.2 – RRI(PositionalIndex), Tensor Similarity et k-NN

Dans cette deuxième exécution, nous avons utilisé la même terminologie que pour 2.1, mais l'espace sémantique a été construit en utilisant RRI sur des indexes positionnels. Le calcul des vecteurs de mots-clés utilise l'opérateur Tensoriel de Semantic Vectors. Les mots-clés retenus pour un document d sont les k plus proches voisins du document d dans la terminologie, k étant le nombre de mots-clés pour le document d.

5.2.3 Run 2.3 – RRI(chunks), BN et règles

Cette exécution est identique à la deuxième exécution de la Piste 1 décrite en 5.1.2, la terminologie obtenu par la méthode décrite en 3.2 à laquelle on ajouté les mots-clés du corpus d'apprentissage elle contenaint 3270 candidats mots-clés.

5.3 Discussion

Nous pouvons voir que les exécutions 1.2 et 2.3 sont celles qui obtiennent les meilleurs résultats, ce qui nous conforte dans nos hypothèses de départ. Les exécutions officielles nous ne permettent pas de comparer les performances des espaces "enrichis" par des chunks et des espaces RRI avec indexes positionnels, nous avons effectué une exécution 2.2bis avec un espace "enrichi" et k-NN le F-score obtenu est de 0.4186, le résultat est sensiblement meilleur que l'exécution 2.2.

Rappelons que pour le 1.3, on a combiné les résultats de 1.1 et 1.2 de en donnant plus de poids aux candidates-termes longues (cette règle n'ayant pas été incluse dans le système de règles décrit en 4.2). Etant donné que le F-score obtenu (0.7486) se trouve au mi-chemin entre le F-score de 1.1 et de 1.2, nous ne pouvons pas réellement conclure quand à la pertinence de cette règle.

Conslusion

Dans cet article, nous avons présenté un système d'attribution de mots-clés à des articles scientifiques, qui se base sur des espaces sémantiques construit en utilisant RRI. Puis nous avons essayé d'améliorer les performances du systèmes par deux moyens : (i) en enrichissant les espaces sémantiques par des informations issues d'une analyse linguistique de surface, et (ii) en définissant une procédure de décision basée sur une combinaison de réseaux bayésiens et de systèmes à base de règles. Les résultats obtenus montrent que ces deux hypothèses se sont révélées payantes et qu'elles améliorent sensiblement les résultats obtenus par une approche RRI seul (qui obtient déjà des résultats honorables).

Références

ABNEY, S. (1991). *Principle-Based Parsing*, chapitre Parsing By Chunks. Kluwer Academic Publishers.

AÏT-KACI, H. et BONNARD, P. (2011). Probabilistic production rules. Rapport technique, IBM.

BARBER, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

BINGHAM, E. et MANNILA, H. (2001). Random projection in dimensionality reduction : Applications to image and text data. *In in Knowledge Discovery and Data Mining*, pages 245–250. ACM Press.

COHEN, T., SCHVANEVELDT, R. et RINDLESCH, T. (2009). Predication-based semantic indexing : Permutations as a means to encode predications in semantic space. *In Proceedings of the AMIA Annual Symposium*, pages 114–118.

COHEN, T., SCHVANEVELDT, R. et WIDDOWS, D. (2010a). Reflective random indexing and indirect inference : A scalable method for the discovery of implicit connections. *Biomed Inform*, 43(2): 240–256.

COHEN, T., WIDDOWS, D., SCHVANEVELDT, R. et RINDLESCH, T. (2010b). Logical leaps and quantum connectives : Forging paths through predication space. *In Proceedings of the AAAI Fall 2010 symposium on Quantum Informatics for cognitive, social and semantic processes (QI-2010)*.

EL GHALI, A. (2011). Expérimentations autour des espaces sémantiques hybrides. *In Actes de l'atelier DEFT'2011*, Montpellier.

HARRIS, Z. (1968). *Mathematical Structures of Language*. John Wiley and Son, New York.

JONES, M. N. et MEWHORT, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1):1–37.

KANERVA, P., KRISTOFERSON, J. et HOLST, A. (2000). Random Indexing of Text Samples for Latent Semantic Analysis. *In* GLEITMAN, L. et JOSH, A., éditeurs : *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah. Lawrence Erlbaum Associates.

LANDAUER, T. K. et DUMAIS, S. T. (1997). A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

LUND, K. et BURGESS, C. (1996). Producing high-dimensional semantic space from lexical co-occurence. *Behavior research methods, instruments & computers*, 28(2):203–208.

RANGAN, V. (2011). Discovery of related terms in a corpus using reflective random indexing. *In Proceedings of Workshop on Setting Standards for Searching Electronically Stored Information In Discovery Proceedings (DESI-4)*.

SAHLGREN, M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Thèse de doctorat, Department of Linguistics Stockholm University.

VEMPALA, S. S. (2004). *The Random Projection Method*, volume 65 de *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society.

WIDDOWS, D. et COHEN, T. (2010). The semantic vectors package : New algorithms and public tools for distributional semantics. *In Proceedings of the Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010)*.