Combiner espaces sémantiques, structure et contraintes

Adil El Ghali[1, 2] Kaoutar El Ghali[2]
(1) European Commission, Joint Research Centre – via Enrico Fermi 2749, 21027 Ispra VA, Italy
(2) LUTIN, Cité des sciences et de l'industrie – 30, avenue Corentin Cariou, 75930 Paris cedex 19
adil.el-ghali@jrc.ec.europa.eu, kaoutar.elghali@lutin-userlab.fr

Résumé. Dans la lignée des méthodes que nous avons présenté lors de nos précédentes participations au DEFT, nous présentons cette année un ensemble de méthodes qui combinent une représentation de la sémantique dans des espaces vectoriels construits avec Random Indexing avec méthode s'appuyant sur une formalisation de la structure des genres poétiques pour la tâche 1 et une approche à base de contraintes pour la tâche 4.

Abstract. In line with the methods we have introduced in our previous participations DEFT, we present this year some methods that combine a representation of the semantic in vector spaces constructed with Random Indexing with a method based on a formalization of the structure of poetic genres for Task 1, and an approach based on constraints for the task 4.

Mots-clés : Espaces sémantiques, Random Indexing, contraintes, structure poétique, clustering.

Keywords: Semantic spaces, Random Indexing, constraints, poetry structure, clustering.

1 Introduction

L'édition de cette du DEFT proposait quatre tâches assez différentes. Nous avons choisi d'en aborder deux (i) la catégorisation du genre d'un texte littéraire (tâche 1), et (ii) l'assignation de session à des articles scientifiques (tâche 4). Notre approche pour les deux tâches a été basé sur le même fondement : une représentation du sens des textes dans des espaces sémantiques construits en utilisant Random Indexing. Il y a toutefois une grande différence entre la façon dont les tâches ont été adressés.

Dans la première (tâche 1), nous avons utilisé une méthode relativement simple pour assigner les catégories "sémantiques" qui consiste à représenter les documents et les catégories dans un seul espace vectoriel qui permet leur comparaison, et de calculer à partir de là les catégories les plus proches pour chacun des documents. Nous avons ensuite calculer séparément la sous-catégorie poétique de chaque document appartenant à cette catégorie en se basant sur une méthode qui tente de modéliser par une série de descripteurs de structure les genres poétiques.

Alors que pour la seconde (tâche 4), nous avons considéré le problème d'assigner une session aux articles d'une conférences scientifiques comme un problème de clustering contraint par le nombre d'articles dans les sessions et par les distances relatives des articles obtenues dans l'espace sémantique représentant les articles de la conférence.

2 Corpus

2.1 Tâche 1

La tâche 1 a pour but de classer un texte littéraire court selon le genre qui lui correspond. Le corpus d'apprentissage est constitué d'œuvres publiées sur *Short Edition*, éditeur en ligne de textes courts. Il est composé de documents $N_{app} = 2328$ répartis en 7 catégories : *autres*, *chronique*, *fantastique - sf*, *jeunesse*, *noir*, *poésie* et *émotions* (figure 1).

Les catégories chronique et émotions sont celles qui sont associées au plus grand nombre de documents dans le corpus ($N_{chronique} = 1055$, $N_{emotions} = 1771$).

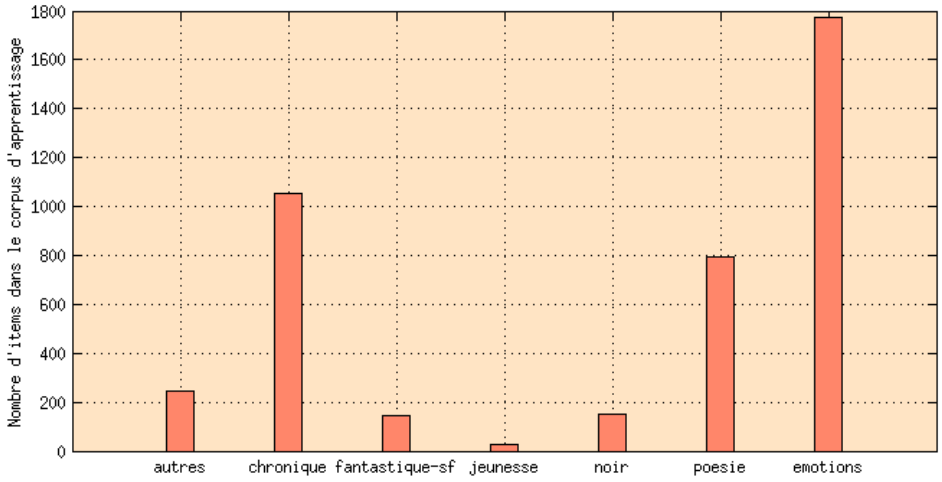

FIGURE 1 – Répartition des documents par catégorie

Chaque document peut être associé à plusieurs catégories. Les intersections entre celles-ci sont représentées dans la figure 2. Les catégories les plus représentées (*chronique* et *émotions*) présentent naturellement les intersections les plus importantes avec les autres catégories. La seule intersection nulle est entre les catégories *jeunesse* et *noir*.

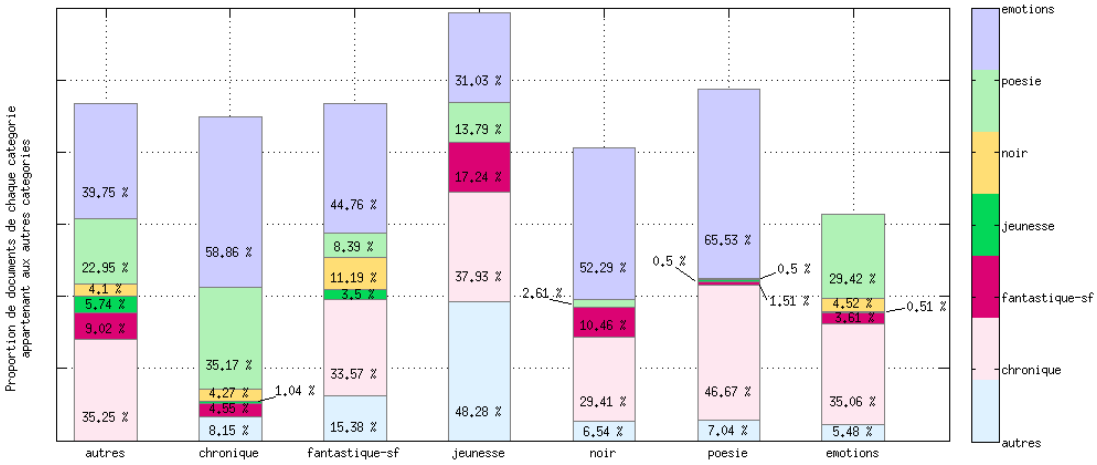

FIGURE 2 – Intersection des catégories dans le corpus d'apprentissage

Chaque catégorie regroupe plusieurs sous-catégories (figure 3) ; celles-ci sont non-spécifiques, à l'exception de celles qui sont poétiques. Aussi, chaque œuvre peut appartenir à, au plus, 5 sous-catégories ; et chaque poème appartient à une seule et unique sous-catégorie poétique. Il n'existe donc pas de fonction entre l'ensemble des documents est celui des sous-catégories non-spécifiques. En revanche, il existe une application surjective entre celui des poèmes et celui des sous-catégories poétiques.

2.2 Tâche 4

La tâche 4 concerne la classification d'articles scientifiques présentés en communication orale lors des dernières conférences TALN. Il s'agit d'identifier, à partir de l'article, son résumé et ses mots-clés, la session dans laquelle il a été présenté.

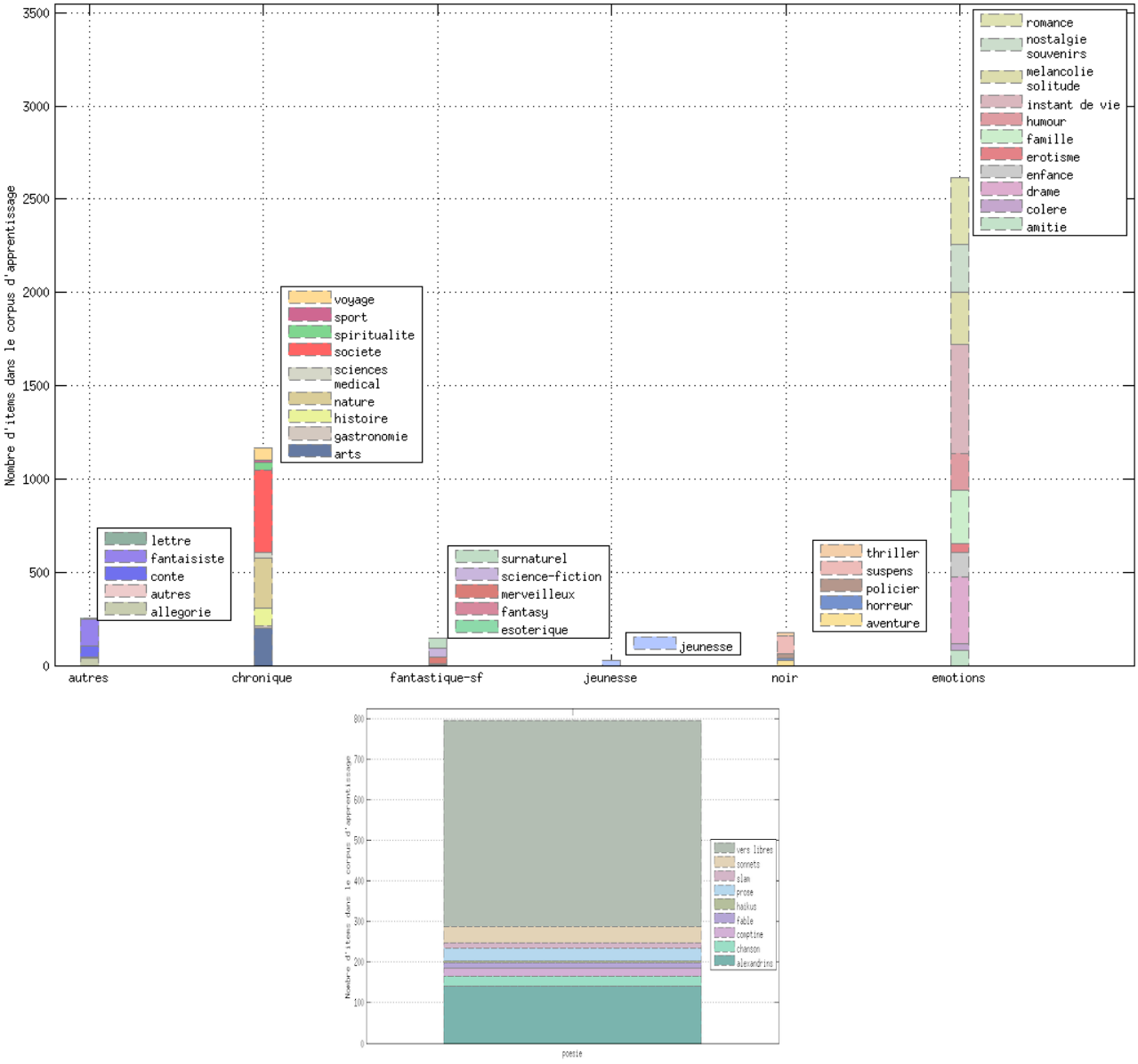

FIGURE 3 – Répartition des documents dans les sous-catégories non-spécifiques

Le corpus d'apprentissage se compose de 208 articles présentés en 2002, 2005 et de 2007 à 2011, et répartis en 43 sessions scientifiques. Le corpus d'apprentissage se compose de 55 articles présentés en en 2012 et 2013, et répartis en 16 sessions. Le nombre d'articles par session est fourni pour le corpus de test.

3 Espaces sémantiques

Les modèles vectoriels de représentation sémantique de documents, par exemple LSA (Landauer & Dumais, 1997), HAL (Lund & Burgess, 1996) et RI (Kanerva *et al.*, 2000), sont des méthodes algébriques représentant les documents et les mots dans des espaces vectoriels, fonction de l'environnement textuel dans lequel ceux-ci apparaissent. Ces modèles permettent de construire un espace sémantique dans lequel les mots sont représentés comme des vecteurs d'un espace vectoriel de grande dimension, où leurs distances les uns aux autres représentent leur similarité sémantique. En effet, en se basant sur l'hypothèse distributionnelle de Harris, qui stipule que les mots apparaissant dans des contextes similaires tendent à avoir un sens similaire, ces méthodes transforment l'analyse distributionnelle d'un corpus en espace sémantique.

Une problématique commune à ces méthodes est la construction d'espace sémantique à partir des matrices mot-document ou mot-mot issues des analyses distributionnelles. L'extraction de concepts se fait par le biais de méthodes mathématiques de réduction de dimensionnalité, permettant ainsi de projeter le corpus dans un espace vectoriel de dimension réduite. Le but principal de ces méthodes mathématiques est de construire un modèle simplifié pertinent rendant compte des variations de fréquence, en décorrélant des données multidimensionnelles et en éliminer les dimensions considérées comme « bruitées ». LSA utilise la décomposition en valeurs singulières (SVD), alors que HAL utilise l'analyse en composantes principales (PCA) ; ces deux méthodes mathématiques sont des outils classiques de factorisation de matrices. Le coût calculatoire ainsi que le peu de malléabilité que présentent ces outils de factorisation compliquent l'utilisation de LSA et HAL, comme modèles sémantiques vectoriels.

Différentes mesures de similarité peuvent être utilisées pour approximer la similarité sémantique. Nous pouvons citer, entre autres, le coefficient de Dice et l'indice de Jaccard. En fouille de texte, on utilise classiquement la mesure cosinus de l'angle entre deux vecteurs, représentant deux mots ou deux groupes de mots.

Nous avons choisi d'utiliser Reflective Random Indexing (Cohen *et al.*, 2010), variante de Random Indexing, comme méthode de construction d'espace sémantique. Le but RI est d'aboutir à une réduction de la dimensionalité sans la complexité calculatoire d'outils de factorisation de matrices. Ainsi, RI ne passe pas par la construction de matrices d'occurrences, mais construit directement l'espace des « concepts » dit espace sémantique ; en se basant sur le lemme de Johnson-Lindenstrauss (Vempala, 2004; Bingham & Mannila, 2001). Ce lemme stipule qu'un ensemble de vecteurs de grande dimension peuvent être projetés orthogonalement dans un sous-espace de dimension réduite par une matrice de projection aléatoire tout en préservant les distances à une petite distorsion près.

Soit $0 < \epsilon < 1$, $y_1, ..., y_N$ un ensemble de vecteurs de $\mathbb{R}^d$ et R une matrice de projection orthogonale telle que ses éléments sont indépendamment et identiquement distribués selon une loi normale centrée réduite.

La construction de l'espace sémantique de dimension par Random Indexing se fait par la mise en œuvre de l'algorithme suivant :

- Créer une matrice A de taille $d \times k$ contenant les vecteurs indexes, où d est le nombre de documents ou de contextes dans le corpus ; ces vecteurs sont creux et identiquement et indépendamment distribués selon une loi normale centrée réduite $[0...0... - 1...0...0...1...0...0]^T$.
- Créer une matrice B de taille $t \times k$ contenant les vecteurs termes, où t est le nombre de termes différents dans le corpus ; ces vecteurs sont initialisés à des valeurs nulles pour débuter 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.

L'aspect « Reflective » dans RRI réside dans le fait que les vecteurs-termes sont re-projetés sur les vecteurs indexes. Puis les trois étapes de l'algorithme, plusieurs fois si besoin de plus de précision, l'on arrive à construire un espace de un espace sémantique qui capture les « patrons » essentiels de co-occurrence du corpus et dans lequel termes et documents sont comparables.

Plusieurs implémentations libre de RI sont disponibles, nous utilisons la librairie Semantic Vectors [1] (Widdows & Cohen, 2010).

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

4 Détermination des catégories d'un texte

Dans la lignée des méthodes que nous avions utilisé pour les précédentes éditions du DEFT (El Ghali *et al.*, 2012; El Ghali & Hoareau, 2010; Hoareau & El Ghali, 2009), l'approche que nous présentons cette année est basée sur l'exploitation des similarités entre documents dans un espace sémantique construit avec RRI. Pour la tâche 1, il s'agissait d'assigner les sous-catégories pour les textes donnés.

La méthode que nous avons utilisé comporte trois étapes :

1. construire à un espace sémantique avec les documents du corpus d'apprentissage et de test ;
2. calculer une signature "représentative" pour chacune des sous-catégories en utilisant les informations fournies dans le corpus d'apprentissage ;
3. assigner les catégories au documents du corpus d'apprentissage en fonction de leurs similarités aux signatures des sous-catégories.

Construction des espaces sémantiques La construction des espaces sémantiques utilise la méthode RRI décrite dans la section précédente. Nous avons utilisé deux configurations d'espace. La première consistait à mettre l'ensemble des documents du corpus d'apprentissage et du corpus de test dans le même espace, alors que dans la deuxième configuration nous avons créé un espace séparé pour chacune des catégories de haut niveau (Poetik, Chronique, Emotions, Fantastique-SF, Jeunesse, Noir, Autres).

Partant de l'hypothèse que des textes dans la même catégories avaient une identité sémantique propre, nous prévoyions que la deuxième configuration serait plus performantes. Ceci étant dit, le nombre relativement bas de textes dans certaines catégories pouvait amené à une chute de performances dans cette configuration.

Calcul de signature des sous-catégories La signature d'une sous-catégorie est une représentation de la sous-catégorie dans l'espace sémantique représentant les documents. Cette signature doit avoir comme propriété principale d'être comparable aux représentations des documents.

Nous définissons comme signature d'une sous-catégorie C, le vecteur $\vec{v}_C$ obtenu en sommant les vecteurs de tous les documents du corpus d'apprentissage appartenant à cette sous-catégories, éventuellement en y associant un poids correspondant à l'importance de la sous-catégorie donnée par le rang de la catégorie. Formellement, le vecteur d'une sous catégorie se définit comme suit :

$$\vec{v}_C = \sum_{d|d \in C} w_d.\vec{v}_d \text{ ; où } w_d \text{ est le poids de la sous-catégorie pour } d$$

Assignation des sous-catégories L'algorithme d'assignation est des sous-catégories se contente de déterminer les sous-catégories les plus proches pour un document d_t du corpus de test, en se limitant aux sous-catégories C_i dont la distance $d(\vec{d_t}, \vec{C_i})$ est inférieure à la distance maximale entre la signature de C_i et les documents appartenant à C_i. Les sous-catégories sont ordonnées en fonction de leur similarité au document d_t.

5 Détermination du genre poétique

5.1 Versification française

Le vers, en versification française, est mesuré (Sorgel, 1986). Le mètre syllabique est le nombre de syllabes comptées dans un vers. Une syllabe est une unité prosodique, c'est la plus petite unité de combinaisons de sons. Une syllabe est constituée de deux éléments : une attaque, et une rime, formée d'un noyau et d'un coda. L'élément primordial d'une syllabe est son noyau. En français, il s'agit obligatoirement d'un élément vocalique. Dans la versification française, plusieurs règles régissent le compte des syllabes : les élisions de « e caduc », les diérèses, et les synérèses.

Élision des « e caduc » Le « e caduc » désigne la voyelle « e » dont la prononciation varie en fonction de l'environnement syntaxique. On l'associe aux graphies « e », « es » et « ent ». L'élision d'un « e caduc » est une forme d'apocope qui consiste à amuïr cette voyelle. Ainsi, un « e caduc » est élidé :
– systématiquement, en fin de vers ;
– s'il est suivi d'une voyelle ou d'un « h » muet, à l'intérieur du vers ;
– s'il est précédé d'une voyelle, à l'intérieur des mots.

Diérèse et synérèse La diérèse d'une diphtongue est la séparation d'une syllabe en deux par vocalisation d'une spirante. La synérèse d'une diphtongue, par opposition, est la prononciation en une seule syllabe de deux sons voyelles. Les diérèses et synérèses dépendent de critères étymologiques. Théoriquement, une diphtongue comptera donc pour une ou deux syllabes selon qu'elle est issue d'une ou deux syllabes latines (cf. exemples en table 3). Toutefois, dans la pratique, diérèses et synérèses tiennent souvent à la licence poétique, ie les considérations métriques, rythmiques, ou esthétiques du poète.

TABLE 1 – Règles de décompte des syllabes dans les diphtongues, *Traité De Prosodie Classique À L'usage Des Classiques Et Des Dissidents*

Diphtongue	Nombre de syllabes	Exemples
ié	2	les mots en i-é-té : so-bri-é-té
oi	1	Toi, roi, voi-là..
oin	1	Loin
io iau	2	Bri-oche mi-au-ler

Rimes La rime est un élément métrique important en poésie. C'est une homophonie entre les phonèmes à la fin d'au moins deux vers.

Les rimes peuvent être continues (AAAA), plates (AABB), croisées (ABAB), embrassées (ABBA), alternées (ABCABC), en rhythmus tripertitus (AABCCB), ou en rhythmus quadripertitus (AAABCCCB).

On appelle rimes féminines celles se terminant par un « e caduc » et rimes masculines les autres (indépendamment du genre du mot). Rimes masculines et féminines ne peuvent rimer ensemble et doivent être alternées en poésie classique.

On appelle rimes pluriel celles finissant par « s », « x », ou « z » et singulier les autres. On ne peut faire rimer une rime singulier et une rime pluriel.

Quelques figures de style en poésie

L'allitération consiste à répéter une ou d'un groupe de consonnes à l'intérieur d'un vers, majoritairement à l'attaque des syllabes accentuées. En français, les consonnes sont classées en cinq familles permettant les allitérations et quelques isolées : les labiales (b, p, f, m, v), les dentales (d, t, l), palatales (j, g, n), les vélaires (k, g, w) et les uvulaires (r).

L'anaphore consiste à commencer un ou un groupe de vers ou de une phrase par le même mot ou le même syntagme.

L'assonance consiste à répéter une voyelle ou un son vocalique dans des mots proches ; plus spécifiquement, il s'agit de répéter dans un vers le dernier son vocalique non caduc à l'intérieur du vers.

La paronomase consiste à rapprocher des paronymes (des mots ayant des graphies ou/et des prononciations proches).

5.2 Détermination automatique des genres poétiques

Dans ce qui suit, nous appelons invariablement poèmes, tous les documents de nature poétique ; strophe, tout bloc marqué par une ligne blanche ; vers, chaque segment représenté par un retour à la ligne ; rime, toute syllabe de fin de ligne.

Nous choisissons d'adopter une stratégie one-vs-all pour attribuer une sous-catégorie poétique à chaque poème. 7 classifieurs binaires $(\mathrm{f}_k)_{k\in[1,7]}$ sont donc construits en utilisant un sous-ensemble de l'ensemble des descripteurs extraits. La

sélection de descripteurs pour chaque classifieur se fait par selon la connaissance experte des différentes sous-catégories poétiques. Classiquement, la prédiction de chaque classifieur est associée à un score de confiance, la classe prédite est celle avec le plus haut score de confiance. L'application de ce type de prise de décision pose problème dans notre cas d'étude. En effet, certaines poèmes peuvent appartenir à plusieurs genres poétiques ; par exemple, une fable peut être écrite en prose. Aussi, la prise de décision dans notre système doit donc rendre compte de ce type de cas. Nous choisissons donc d'établir un système de règles pour la prise de décision encodant les interactions entre les différentes sous-catégories poétiques, qui se base sur une connaissance experte du domaine.

5.2.1 Extraction des descripteurs

Pour extraire les descripteurs des poèmes, nous procédons, en premier lieu à un étiquettage morpho-syntaxique des documents. Cela a pour visée de lever l'ambiguïté de certaines terminaisons ; par exemple, la terminaison « ent » dans un verbe est associée à un « e », alors qu'elle est associé à un « en » dans les adverbes. Nous procédons, ensuite, à une phonétisation des poèmes en adoptant un respect strict de la règle d'élision du « e caduc » et des règles de décompte des syllabes dans les diphtongues.

FIGURE 4 – Extraction des descripteurs

Chaque poème d_i est représenté par l'ensemble des descripteurs suivants $F_i^{(j)}$:
– S_i, le nombre de strophes,
– $[V_i^{(1)}, ...V_i^{(S_i)}]$, les nombres de vers pour chaque strophe,
– $[M_i^{(1)}, ...M_i^{(\sum V_i)}]$, le mètre (nombre de syllabes dans chaque vers),
– $[Me_i^{(1)}, ...Me_i^{(S_i)}]$, les métriques moyennes par strophe,
– I_i , l'isométrie (1 si tous les vers ont le même mètre, 0 sinon),
– $[R_i^{(1)}, ...R_i^{(\sum V_i^{(j)})}]$, les rimes,
– $[St_i^{(j)}]$, les indices des strophes répétées dans le poème,
– Al_i, la proportion de d'allitérations,
– An_i, la proportion de d'anaphores,
– As_i , la proportion de d'assonances,
– Pa_i , la proportion de paranomases dans le poème,
– W_i, le nombre moyen de mots avant chaque retour à la ligne,
– P_i, la proportion de retour à la ligne finissant par une marque de ponctuation,
– Rr_i, la proportion de règles de rimes non-respectées dans le poème.

5.2.2 Classifieurs binaires

Classifieur Sonnet Le sonnet est un poème isométrique de quatorze vers, composée de deux quatrains (pièce de quatre vers) suivis de deux tercets (pièce de trois vers). La disposition des rimes est soumise à des règle fixes :
– les deux quatrains sont construits sur le même modèle et sur les mêmes rimes, généralement embrassées (ABBA), plus rarement croisées (ABAB) ;
– le sizain comporte un distique sur une rime et un quatrain aux rimes croisées (CCD EDE) ou embrassées (CCD EED).
Les sonnets irréguliers représentent des altérations de la disposition des quatrains et des tercets (table 2), ou du nombre de rimes : les deux quatrains construit sur 4 rimes au lieu de 2, ou de la métrique (sonnet hétérométrique, sonnet Layé).

Dénomination du type de sonnet	Disposition (Q : Quatrain, T : tercet)
Sonnet à rebours	T/T // Q/Q
Sonnet polaire	Q // T/T // Q
Sonnet alterné	Q/T/Q/T
Sonnet quinzain	Q/Q // T/T // Monostique

TABLE 2 – Formes des sonnets

Pour classer un poème d_i dans la catégorie Sonnet, nous utilisons un arbre de décision, prenant en considération le nombre de strophes S_i, le nombre de vers dans chaque strophe $[V_i^{(1)}, ...V_i^{(S_i)}]$, la disposition des rimes dans chaque strophe $[R_i^{(1)}, ...R_i^{(\sum V_i^{(j)})}]$ (figure 5). Cela permet de représenter toutes les formes de sonnets réguliers et irréguliers listées ci-dessus. Un poème est classé comme Sonnet s'il remplit toutes les conditions suivantes :

– le nombre de strophes est 4 ou 5
– le nombre de vers par strophes correspond à une des formes de la table 2 *ie* [4, 4, 3, 3], [3, 3, 4, 4], [4, 3, 4, 3], [4, 3, 3, 4], [4, 4, 3, 3, 1]
– les rimes de chaque quatrain correspondent au schéma ABBA ou ABAB
– les rimes des tercets correspondent au schéma CCD EDE ou CCD EED ; à noter : dans les sonnets à rebours, les tercets sont retournés, on applique donc un miroir sur le vecteur des rimes.

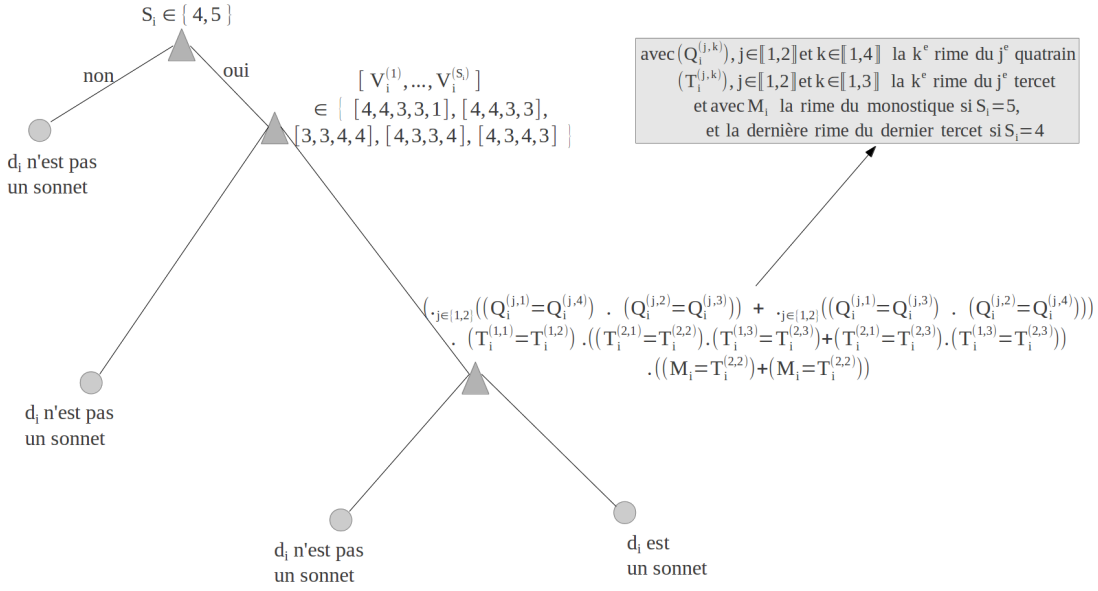

FIGURE 5 – Arbre de décision pour la Classification de Sonnet

Classifieur Haïku Le haïku est un poème d'origine japonaise comportant traditionnellement 17 mores en trois segments 5-7-5. En français, c'est un poème composé de tercets formés d'un heptasyllabe encadrés de deux pentasyllabes. Pour classer un poème d_i dans la catégorie Haïku, nous utilisons un arbre de décision, prenant en considération le nombre de strophes S_i, le nombre de vers dans chaque strophe $[V_i^{(1)}, ...V_i^{(S_i)}]$, le mètre de chaque vers $[M_i^{(1)}, ...M_i^{(\sum V_i)}]$ (figure 6). Un poème est classé comme Haïku si :

– $[V_i^{(1)}, ...V_i^{(S_i)}]$ est constant et égal à 3,
– $M_i^{(1)} = \begin{cases} 7 & \text{si } j = 2 \\ 5 & \text{sinon.} \end{cases}$

Classifieur Alexandrin L'alexandrin est un vers formé de deux hémistiches de six syllabes chacun, s'articulant à la césure. L'alexandrin classique présente une césure centrale fixe correspondant à une pause grammaticale ; et précédée d'une

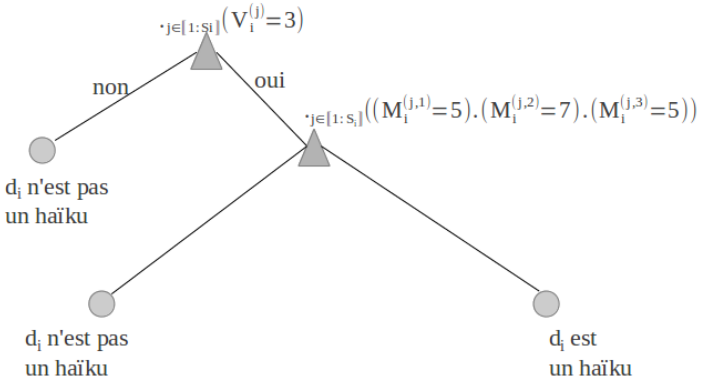

FIGURE 6 – Arbre de décision pour la Classification de Haïku

voyelle tonique, et ne tolère donc pas d'être précédée d'une syllabe féminine surnuméraire. Ces règles sont, toutefois, affaiblies dans l'alexandrin romantique ; des césures illicites dans le modèle classique peuvent ainsi être acceptées.

Pour classer un poème d_i dans la catégorie Alexandrin, nous utilisons un arbre de décision, prenant en considération le nombre de vers dans chaque strophe $[V_i^{(1)}, ...V_i^{(S_i)}]$, le mètre de chaque vers $[M_i^{(1)}, ...M_i^{(\sum V_i)}]$. Un poème est classé comme Alexandrin, si :
– Rs_i est égale à 1,
– $[M_i^{(1)}, ...M_i^{(\sum V_i)}]$ est constant et égal à 12.

Classifieur Prose Les poèmes en prose n'ont pas la forme d'un poème, ie qu'ils ne sont découpés ni en strophes, ni en vers. Ils se caractérisent par une recherche de rythme dans les phrases. Il n'y a pas de rimes mais une recherche d'écho sonore avec allitération, assonance et rimes intérieures. Pour classer un poème d_i dans la catégorie Prose, nous utilisons un classifieur bayésien naïf, prenant en considération le nombre moyen de mots avant chaque retour à la ligne W_i, la proportion de retour à la ligne finissant par une marque de ponctuation P_i. Il s'agit, en fait, d'estimer les paramètres des densités de probabilités de chaque descripteur à partir des données du corpus d'apprentissage et de calculer le log de vraisemblance d'appartenance à la sous-catégorie Prose a posteriori des poèmes du corpus de test :

$$ln\frac{p(Prose/d_i)}{p(\neg Prose/d_i)} = ln\frac{p(Prose)}{p(\neg Prose)}\sum_{\lambda\in\Lambda} ln\frac{p(\lambda/Prose)}{p(\lambda/\neg Prose)}, \text{ où : } \Lambda = \{W, P, Rr\}$$

Vers libres La poésie libre ne répond pas à une structure régulière ; n'obéissant ni à la métrie, ni à la régularité des strophes, ni aux règles concernant les rimes. Elle cherche néanmoins une cohérence rythmique.

Pour classer un document d_i dans la catégorie Vers libres, nous prenons en considération la régularité des strophes $\frac{1}{S_i}\sum_{j=1}^{S_i} {V_i^{(j)}}^2 - (\sum_{j=1}^{S_i} V_i^{(j)})^2$, la variabilité du mètre $\frac{1}{\sum V_i^{(j)}}\sum_{j=1}^{S_i} {V_i^{(j)}}^2 - (\sum_{j=1}^{S_i} M_i^{(j)})^2$, la proportion de règles de rimes non-respectées dans le poème R_r.

Fable Une fable est un court récit en vers ou occasionnellement en prose qui vise à donner de façon plaisante une leçon de vie. Elle se caractérise souvent par la mise en scène d'animaux qui parlent mais peut également mettre en scène d'autres entités ou des êtres humains. Elle a pour but d'exprimer une morale à la fin ou au début de la fable quand elle n'est pas implicite.

Pour classer un document d_i dans la catégorie Fable, nous utilisons un classifieur bayésien naïf, prenant en considération la longueur du poème $\sum_{i=1}^{S_i} V_i^{(j)}$, la proportion de règles de rimes non-respectées dans le poème R_r, l'isométrie I_i.

Chanson La chanson est un poème à chanter composé de stances égales appelées couplets, séparées généralement par un leitmotiv, le refrain.

Pour classer un document d_i dans la catégorie Chanson, nous utilisons un classifieur bayésien naïf, prenant en considération la proportion de strophes répétées $\frac{|(St_i^{(j)})|}{S_i}$ et la régularité des strophes non-répétées :

Comptine Les comptines sont des textes à dire ou à chanter ; elles sont caractérisées par de courtes séquences à construction rythmée.

Pour classer un document dans la catégorie Comptine, nous utilisons un classifieur bayésien naïf, prenant en considération la longueur du poème $\sum_{j=1}^{S_i} V_i^{(j)}$, et les moments d'ordre 0 et 1 du mètre $[M_i^{(1)}, ...M_i^{(\sum V_i)}]$.

Slam Le slam est une forme de poésie orale. Apparentés au « *spoken word* », les slams sont des textes destinés à être lus, essentiellement scandés ; jouant avec l'harmonie imitative, avec allitérations, assonances, et paronomases.

Pour classer un document d_i dans la catégorie Slam, nous utilisons un classifieur bayésien naïf, prenant en considération la proportion d'allitérations Al_i, la proportion d'anaphores An_i, la proportion d'assonances As_i, la proportion de paranomases Pa_i.

6 Assigner la session à un article

La tâche 4 avait pour but d'assigner à chaque article du corpus de test la session à laquelle il était présenté. Nous disposions des textes des articles, de leurs mots clés ainsi que du nombre d'articles par session.

Au lieu de considérer cette tâche comme une tâche de catégorisation, nous avons opté pour une méthode légèrement différente : la considérer comme une tâche de clustering contraint s'inspirant de (Wagstaff *et al.*, 2001) et de (Bilenko *et al.*, 2004). Nous avons donc implanté une variante de COP-K-means, dans laquelle nous avons intégrés la contrainte du nombre de d'articles par session.

K-Means est un algorithme de partitionnement de données basé sur la construction d'une partition de Voronoï de taille K générée par les moyennes et la mise à jour des barycentres de chaque cluster. Il s'agit, en fait, d'une optimisation locale de la somme des moindres carrés entre chaque point et le barycentre du cluster auquel il appartient. Une classification par K-means sous contraintes binaires peut être formulée sous forme d'une optimisation multi-objectif. En effet, si l'on associe un coût à la violation de chacune des contraintes binaires, l'algorithme COP-K-means peut être traduit comme une optimisation locale de moindres carrés régularisés.

Étant donné un ensemble de documents. L'algorithme se présente comme suit :
– Initialiser les barycentres des clusters
– Répéter jusqu'à convergence ;
  – assigner à chaque document son cluster le plus proche qui minimise le coût de violation des contraintes ;
  – mettre à jour le barycentre de chaque cluster ;

L'initialisation des clusters a été effectué en utilisant la proximité dans l'espace sémantique entre les termes des noms de sessions et les articles. Et nous avons utilisé deux types de contraintes : (i) le nombre d'articles par session ; (ii) les distances entre documents comme coût de violation des contraintes "must-link", "cannot-link".

7 Détails des soumissions

Nous avons soumis trois exécutions, une pour la tâche 4 et deux pour la tâche 1 dans lesquelles nous avons fait varier les caractéristiques des espaces sémantiques et le calcul de signature.

Ces exécutions sont résumées dans le tableau ci-dessous :

TABLE 3 – Détail des soumissions

ID	score	Détail de la soumission
19-1-1	0.4278	Espace sémantique commun à toutes les catégories de haut-niveau ; signature de sous-catégorie non pondérée ($w_i = 1$)
19-1-2	0.2599	Espace sémantique séparé pour chacune des catégories de haut-niveau ; signature de sous-catégorie pondérée par le rang
19-4-1	1	Espace sémantique par édition ;

8 Conclusions

Dans cette édition du DEFT'14, nous avons abordé les tâches 1 et 4 en utilisant comme base des systèmes que nous avions développé pour les précédentes éditions du DEFT. Dans ces méthodes, l'élément central est la représentation des documents dans des espaces sémantiques de grande dimensions, qui permet de comparer les documents entre eux, mais aussi d'abstraire des représentations pour des catégories qui sont elle-mêmes comparables aux documents.

La particularité des tâches de cette édition nous a permis d'enrichir notre panoplie avec une nouvelle méthode de clustering sous contraintes et une approche pour la formalisation des genres poétiques.

Références

BILENKO M., BASU S. & MOONEY R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of 21st International Conference on Machine Learning (ICML-2004)*, p. 81–88, Banff, Canada.

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

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

EL GHALI A. & HOAREAU Y. V. (2010). μ-alida : expérimentations autour de la catégorisation multi-classes basée sur alida. In *Actes de l'atelier DEFT'2010*, Montreal, Canada.

EL GHALI A., HROMADA D. & EL GHALI K. (2012). Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés. In *JEP-TALN-RECITAL 2012, Atelier DEFT 2012 : DÉfi Fouille de Textes*, p. 77–90, Grenoble, France : ATALA/AFCP.

HOAREAU Y. V. & EL GHALI A. (2009). Approche multi-traces et catégorisation de textes avec Random Indexing. In *Dans les actes de atelier de clôture de l'édition 2009 du défi fouille de texte*, Paris, France.

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

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

SORGEL G. (1986). *Traité de prosodie classique à l'usage des classiques et des dissidents*. La nouvelle proue. Association "Les Amis de Marcel Chabot.

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

WAGSTAFF K., CARDIE C., ROGERS S. & SCHROEDL S. (2001). Constrained k-means clustering with background knowledge. In *In ICML*, p. 577–584 : Morgan Kaufmann.

WIDDOWS D. & 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)*.