13éme atelier sur la Fouille de Données Complexes (FDC)
Extraction et Gestion des Connaissances (EGC 2016)

Organisateurs :

- Cécile Favre (ERIC, Université Lyon 2)
- Germain Forestier (MIPS, Université de Haute-Alsace)
- Camille Kurtz (LIPADE, Université Paris Descartes)

Le groupe de travail "Fouille de Données Complexes"

La trèzième édition de l'atelier sur la fouille de données complexes est organisée par le groupe de travail EGC "Fouille de Données Complexes". Ce groupe de travail rassemble une communauté de chercheurs et d'industriels désireux de partager leurs expériences et problématiques dans le domaine de la fouille de données complexes telles que le sont les données non-structurées (ou faiblement), les données obtenues à partir de plusieurs sources d'information ou plus généralement les données spécifiques à certains domaines d'application et nécessitant un processus d'extraction de connaissance sortant des itinéraires usuels de traitement.

Les activités du groupe de travail s'articulent autour de trois champs d'action progressifs :

- l'organisation de journées scientifiques une fois par an (vers le mois de juin) où sont présentés des travaux en cours ou plus simplement des problématiques ouvertes et pendant lesquelles une large place est faite aux doctorants ;
- l'organisation de l'atelier "Fouille de Données Complexes" associé à la conférence EGC qui offre une tribune d'expression pour des travaux plus avancés et sélectionnés sur la base d'articles scientifiques par un comité de relecture constitué pour l'occasion ;
- la préparation de numéros spéciaux de revue nationale, dans lesquels pourront être publiées les études abouties présentées dans un format long et évaluées plus en profondeur par un comité scientifique.

Contenu scientifique de l'atelier

Nous avons reçu cette année 9 propositions originales, chacune d'elle a été relue par au moins deux évaluateurs. Pour la majorité des propositions nous avons été en mesure de proposer trois rapports d'experts afin d'offrir un processus scientifique constructif aux auteurs. Vu la qualité des soumissions et l'intérêt des discussions qu'elles pouvaient susciter au sein de l'atelier, nous avons retenu cette année 8 des propositions.

Les articles qui vous sont proposés cette année dans les actes qui suivent explorent une grande variété de complexités, aussi bien dans les données que dans les processus de fouille envisagés.

Remerciements

Les responsables de l'atelier souhaitent remercier vivement toutes les personnes ayant contribué à la tenue de cet atelier. En particulier :

- les auteurs pour la qualité de leurs contributions constituant la base essentielle de discussions fructueuses ;
- les membres du comité de programme et plus généralement tous les relecteurs de cet atelier dont le travail d'évaluation était crucial pour assurer la qualité de l'atelier ;
- les organisateurs d'EGC 2016 qui ont mis en place l'environnement et les moyens pour la réussite des ateliers.

Nous remercions enfin vivement les présidents, Bruno Crémilleux le président du comité de programme et Cyril de Runz le président du comité d'organisation d'EGC 2016 ainsi que les organisateurs des ateliers Amine Ait-Younes et Eric Desjardin.

Cécile Favre	Germain Forestier	Camille Kurtz
ERIC	MIPS	LIPADE
Université Lyon 2	Université de Haute-Alsace	Université Paris Descartes

Membres du comité de lecture

Le Comité de lecture est constitué de :

- Hanane Azzag (LIPN, Univ. Paris 13)
- Nicolas Béchet (IRISA, Univ. Bretagne Sud)
- Alexandre Blansché (Univ. Lorraine)
- Omar Boussaid (ERIC, Univ. Lyon 2)
- Guillaume Cleuziou (LIFO, Univ. Orléans)
- Sami Faiz (INSAT, Université de 7 Novembre de Carthage)
- Cécile Favre (ERIC, Univ. Lyon 2)
- Germain Forestier (MIPS, Univ. Haute Alsace)
- Pierre Gançarski (iCube, Univ. Strasbourg)
- Mehdi Kaytoue (INSA Lyon)
- Camille Kurtz (LIPADE, Univ. Paris 5)
- Mustapha Lebbah (LIPN, Univ. Paris 13)
- Mariem Mahfoudh (LORIA)
- Arnaud Martin (IRISA, Univ. Rennes 1)
- Florent Masséglia (AxIS-Inria Sophia Antipolis)
- Yoann Pitarch (IRIT, Univ. Toulouse 3)
- Mathieu Roche (UMR TETIS, CIRAD)
- Cyril De Runz (CReSTIC, Univ. Reims)
- Cédric Wemmert (iCube, Univ. Strasbourg)

TABLE DES MATIÈRES

Session Audio

RMIS: Système d'extraction de connaissances à partir de données rythmo-mélodiques
Nédra Mellouli, Aichetou Mohamed, Estelle Amy de la Breteque, Victor Stoichita, Herman Akdag 1

Du fichier audio à l'intonation en Français : Graphes pour l'apprentissage de 3 classes intonatives
Martine Cadot, Anne Bonneau 13

Session Apprentissage au cours du temps

Caractérisation d'instances d'apprentissage pour un méta-mining évolutionnaire
William Raynaut, Chantal Soule-Dupuy, Nathalie Valles-Parlangeau, Cedric Dray, Philippe Valet 25

Apprentissage du signal prix de l'électricité : prédictions à long terme
Louis-Victor Pasquier, Antoine Cornuejols, Suzanne Pinson 31

Représentation de données temporelles par un modèle à échelle logarithmique
Brieuc Conan-Guez, Alain Gély, Lydia Boudjeloud-Assala, Alexandre Blansché, Dario Compagno, Arnaud Mercier 43

Session Sémantique

Formation semi-automatique de classes sémantiques couvrantes pour enrichir une ontologie de domaine
Olena Orobinska, Jean-Hugues Chauchat, Natalia Sharonova 51

Mise au point d'un agrégateur de tweets par apprentissage semi supervisé d'ontologies juridiques
Pierre Saurel , Yann Girard, Lambert Pepin 63

Identification de Relations Sémantiques Supportée par des Algorithmes d'Apprentissage Supervisé Utilisant des Mesures de Proximité Sémantique
Jean Petit, Jean-Charles Risch 73

RMIS: Système d'extraction de connaissances à partir de données rythmo-mélodiques

Nédra Mellouli*, Aichetou Mohamed*
Estelle Amy de la Breteque **, Victor Stoichita**
Herman Akdag *

* 2, rue de la Liberté 93526 Saint-Denis Cedex - France
n.mellouli@iut.univ-paris8.fr,
http://www.ai.univ-paris8.fr/
**Centre de Recherche en Ethnomusicologie - LESC - UMR 7186 - CNRS
M.A.E. - Université Paris Ouest - Nanterre La Défense
21 Allée de l'Université - 92023 Nanterre Cedex
victor.stoichita@cnrs.fr
http://crem-cnrs.fr/

Résumé. Le Gamelan Javanais est un orchestre essentiellement constitué de métallophones. Chaque gamelan possède son propre accord, qui caractérise son identité sonore par rapport aux autres gamelans de la même catégorie, comme une signature sonore de la communauté possédant l'instrument. La musique du gamelan se transmet par apprentissage et peu par écrit. Ce sont des aide-mémoires incomplètes mais riches en information. C'est dans ce cadre que s'inscrit notre travail et tente d'analyser un corpus quasi-exhaustif des pièces du gamelan Javanais en proposant deux stratégies croisées d'extraction de descripteurs pertinents.

1 Introduction

Bien loin avant les écrits, les images ou les vidéos, les signaux sonores sont le premier véhicule de l'information et de la communication utilisé par les humains. Si les humains participent à la création des signaux sonores, par l'intermédiaire de la parole, du chant ou de la musique, ils sont aussi distingués par leur remarquable capacité de perception pour décoder certains signaux, parfois même très bruités, comme cela pourrait être le cas de certaines musiques ethniques. Ce lien complexe entre l'homme et la musique a motivé les recherches en Ethnomusicologie qui se consacrent à l'étude des pratiques et des savoirs musicaux, conçus autant comme processus de différenciation socio-culturelle que comme formes d'expression communes à l'homme. Pour la compréhension du savoir musical, certains travaux s'appuient sur la fouille de données musicales. Cette discipline assez récente a été vite bouleversée par la constitution de grandes bases de données, l'intelligence artificielle et la fouille de données, en proposant des techniques de plus en plus pertinentes et efficaces. Différentes problématiques ont été abordées dans la littérature où des outils d'analyse musicale ont été explorés. Certaines

s'intéressent à l'identification des segments parlés et chantés(El Khoury et al., 2008),(Guerid et al., 2012),(Fontan et al., 2015). D'autres études se sont intéressées à la reconnaissance des instruments joués (Souviraà-Labastie et al., 2015),(Razik et al., 2003), ou encore à la reconnaissance de la signature d'un énoncé musical. L'utilisation de la partition engage des connaissances implicites comme le mode de changement des structures. Or une partition même si elle est fortement présente dans toutes les musiques occidentales, elle est loin d'être une représentation universelle. En particulier, certaines musiques ethniques utilisent parfois d'autres systèmes de représentation voire pas de représentation du tout. C'est bien le cas de la musique du Gamelan Javanais qui se transmet par apprentissage et la pratique, compte tenu sa complexité. Le Gamelan Javanais est un instrument collectif où la musique peut se décliner en deux catégories. Certaines peuvent être cycliques et symétriques alors que d'autres peuvent être plutôt linéaires asymétriques et additives. Il existe seulement des aide-mémoires permettant de transcrire les oeuvres sous forme de partitions formelles et incomplètes [1]. Dans cette optique, le projet RMIS [2] porte le développement d'un système d'aide à l'identification des cellules rythmo-mélodiques à partir de leur récurrence dans un énoncé musical ou dans plusieurs énoncés musicaux comparés. L'outil informatique permettra de dépasser la perception subjective d'un ethnomusicologue lorsqu'il cherche à identifier des cellules minimales à l'intérieur d'une musique qu'il étudie. Par conséquent l'outil informatique devra identifier dans un premier temps le plus grand dénominateur commun rythmo-mélodique apparait plusieurs fois à l'intérieur d'un énoncé musical. Dans le cadre de cette étude, plusieurs paramètres devront être pris en compte pour la caractérisation des cellules récurrentes. La notion d'intervalle étant dans certaines cultures musicales très relative et variable, on prendra plutôt en compte la direction (aiguë/haut ou grave/bas) prise par la mélodie. Ce paramètre revient à tracer un chemin, une forme ou un contour mélodique. Dans chaque énoncé musical analysé, nous cherchons à identifier l'ensemble des hauteurs récurrentes ou des "degrés", correspondant chacun à une fréquence plus ou moins précise. L'ensemble des degrés identifiés structurés dans un ordre ascendant ou descendant permet d'obtenir l'échelle mélodique de l'énoncé. Dans un premier temps Nous avons choisi des énoncés non polyphoniques, issus d'un, voire deux instruments. D'un point de vue uniquement rythmique, ce procédé permettra de déceler éventuellement la présence de cycles qui peuvent être longs, dans le jeu de percussions par exemple (ce que l'oreille du musicologue ne parvient pas forcément à l'identifier). Le challenge sera de prendre en compte des variations d'interprétation à l'intérieur de ces cycles, autrement dit d'identifier des cycles même s'ils sont joués avec des variations plus ou moins importantes. La suite de cet article est organisée comme suit. En section 2, nous présentons la méthodologie proposée pour la détection et la reconnaissance des cellules. Les validations expérimentales de cette méthodologie sont présentées dans la section 3. Enfin, nous discutons et nous concluons en section 4.

1. Le site de la cité de la musique à Paris, http ://www.citedelamusique.fr/gamelan/ donne des détails intéressant sur cette musique et l'instrument du Gamelan en général.

2. Projet de recherche financé par la COMUE Université Paris Lumière

2 Méthodologie

2.1 Modélisation des énoncés musicaux incomplets

Le Gamelan Javanais est un ensemble d'instruments de musiques qui opèrent en couches dans une pièce musicale. Les couches reflètent les différentes fonctionnalités attribuées à un groupe d'instruments. Certains instruments déterminent la couche Colotomique correspondant à la structure de la pièce. Cette structure est entièrement définie par les moments où les percussions à gongs sont jouées. Ceci permet de définir le contour mélodique de la pièce. D'autres instruments, souvent les instruments à lames, définissent la mélodie de la pièce (notée par Balungan dans les énoncés du Gamelan). Enfin, des instruments d'élaboration qui permettent de cadrer le rythme (le tempo) de la pièce jouée en fonction de sa Colotomie (s'il s'agit du début de la pièce, milieu ou proche de la fin) et de son espace temporel (pièce jouée le matin, l'après-midi ou le soir, etc.). Dans un énoncé musical, c'est souvent le Balungan qui est noté avec des indications sur les moments où les instruments colotomiques sont joués. Les notes de la Colo-

colotomie	+		+	⌢		+	⌣	+	⌢		+	⌣	+	⌢		+	⌣	+	( )
Balungan	.	2	6	.		2	6	.	2		6	.	2	6		3	2	1	6

TAB. 1 – *Exemple d'un énoncé musical du Gamelan*

tomie sont représentées par des symboles alors que les notes du Balungan sont représentées par des valeurs numériques restreintes. Cette restriction varie en fonction de l'échelle des notes. Il existe deux échelles. Le Slendro est une échelle à 5 notes où l'ensemble des valeurs sont $\{1, 2, 3, 5, 6\}$ et le Pelog une échelle à 7 notes dans l'intervalle $[1..7]$. Quel que soit l'échelle, la structure est décomposée en un nombre variables de Cycles et où chaque cycle est défini par un Gatra (4 temps). Le Gatra est l'unité de mesure minimale d'une pièce de Gamelan. Dans l'écriture javanaise, une pièce est découpée en plusieurs sections. Elle débute obligatoirement par sa description : titre, structure, l'échelle et la restriction. Puis une section optionnelle d'introduction peut avoir lieu. Enfin plusieurs sections constituant le corps de la pièce se succèdent en écriture par pair <Colotomie, Balungan> (voir Tableau 1) composées d'un ou plusieurs cycles. Un cycle est lui-même composé d'un nombre variable de Gatra.

Notations

Une pièce est composée de cycles, chaque cycle contient un nombre n de Gatra. Chaque Gatra est composé de 4 temps. A un indice j (allant de 1 à 4) d'un Gatra i, est associé une note jouée ou de silence $t_{i,j}$ entre 0 et 7 en fonction de l'échelle (la valeur 0 est attribuée au silence). Un Cycle peut être noté comme suit :

$$Cycle_k = G_1 \overbrace{t_{1,1}\ t_{1,2}\ t_{1,3}\ t_{1,4}} \quad G_2 \overbrace{t_{2,1}\ t_{2,2}\ t_{2,3}\ t_{2,4}} \quad \ldots \quad G_n \overbrace{t_{n,1}\ t_{n,2}\ t_{n,3}\ t_{n,4}}$$

Certaines positions dans un Gatra sont très importantes pour caractériser une pièce, en particulier la dernière position (la dernière note). Le dernier Gatra dans un cycle ainsi que la dernière note de celui-ci sont également considérés significatifs pour la caractérisation de la pièce. On

appelle note stable dans un cycle ou dans un Gatra, quand elle est fréquente à la même position. Les valeurs prises par ces Gatras à des positions clés donnent lieu à trois classes rythmo-mélodiques ou pathets par échelle. Un pathet peut être vu comme une restriction des notes jouées aux positions associées aux Gatras dominants et aux positions dominantes dans les Gatras. Nous distinguons les 6 classes suivantes dans les deux échelles réunies :

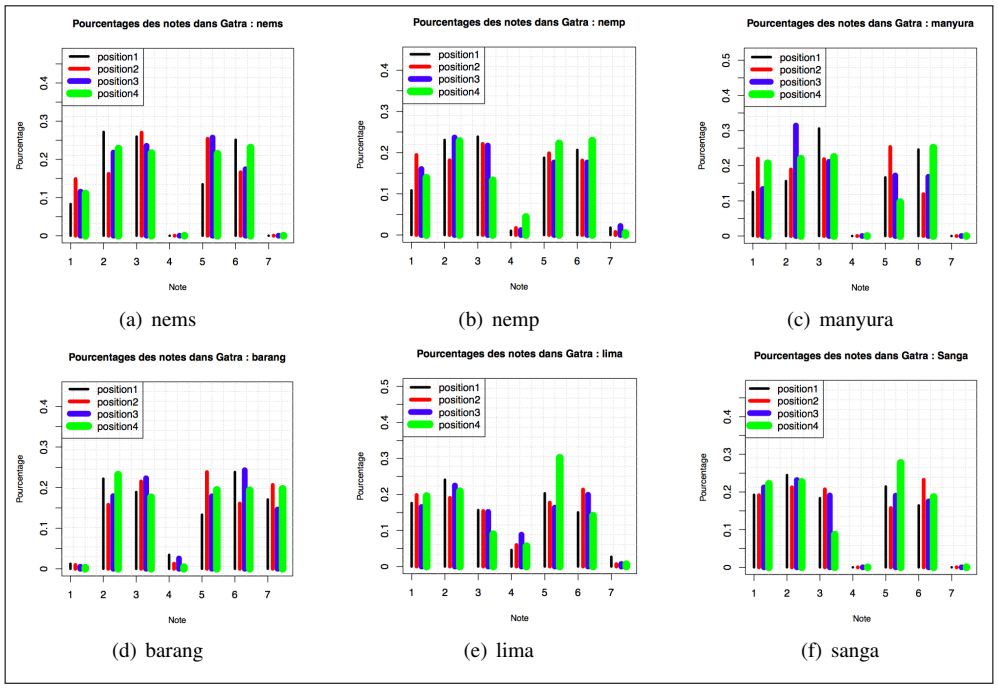
1. Pathet en slendro définit des restrictions sur les notes les plus stables :
   — Manyura : les notes 6 et 2 sont les plus utilisées et récurrentes aux positions les plus importantes (souvent fins des cycles). Note 5 est la moins stable ;
   — Sanga : les notes 5 et 1 sont les plus utilisées, récurrentes aux positions les plus importantes (fin de Gatra). La note 3 est la moins stable ;
   — Nem : les notes 6 et 2 sont les plus utilisées mais la note 3 est la moins stable (elle est utilisée mais change de positions).
2. Pathet en pelog définit une restriction sur l'ensemble des notes utilisé :
   — Barang : l'ensemble $\{7, 6, 5, 3, 2\}$ est utilisé. Absence de 1 et du 4 ;
   — Lima : l'ensemble $\{6, 5, 3, 2, 1\}$ est utilisé. Pas d'utilisation de 7 ou 4, bien que 4 peut être utilisé occasionnellement comme alternatif pour 3 ;
   — Nem : l'ensemble $\{6, 5, 3, 2, 1\}$ est utilisé. Pas d'utilisation de 7 ou de 4, bien que le 4 peut être utilisé occasionnellement comme alternatif pour le 5.

De nombreux systèmes d'indexation de contenu et d'identification des musiques tels que le logiciel Shazam existent dans la littérature. Ils offrent la possibilité de retrouver avec précision une piste musicale à partir d'un extrait, via un téléphone mobile. En ce qui concerne les logiciels dédiés au Gamelan, nous avons pris connaissance de l'architecture du *Gamelan mécanique* réalisée par la cité de la musique et d'une autre application pédagogique VJG (Virtual Javanese Gamelan), disponible sur Google Play pour les systèmes Android [3]. L'application VJG est une plateforme contenant une base très sommaire de partitions écrites et sonores de *Gamelan*. Elle permet également à l'utilisateur d'apprendre à jouer certains instruments du *Gamelan*. Par opposition à ces systèmes, notre problématique se résume à l'ensemble de des questions suivantes : (1) Est-il possible de reconnaitre le Pathet, quel que soit la pièce ? (2) Quels sont les descripteurs les plus représentatifs des partitions ? (3) Quelles sont les méthodes qui permettent une meilleure classification du Pathet ? (4) Quelles sont les partitions et éventuellement les Pathets qui se ressemblent le plus ? Pour y répondre, nous proposons dans la section suivante de construire deux catégories de descripteurs.

2.2 Construction de descripteurs Mélodiques et rythmiques

Selon les experts du Gamelan, le Gatra constitue l'unité minimale de représentation d'une pièce. Les quatre notes que constitue un Gatra peuvent être modélisées de plusieurs manières. Dans notre approche, nous proposons d'étudier deux modèles possibles. Le premier modèle considère un Gatra comme un mot mélodique de 4 caractères. Cette description symbolique de la mélodie a l'avantage de représenter simplement une pièce comme une suite de mots. Un mot est généré par l'expression régulière $[0 - 7]^{+4}$, dont la valeur 0 représente la note de silence. Par conséquent, la recherche de motifs similaires dans une même liste ou bien dans des listes différentes, revient à appareiller les mots deux à deux. Quant au second modèle, il considère un Gatra comme une liste de valeurs. Il en résulte quatre descripteurs quantitatifs différents :

3. https ://play.google.com/store/apps



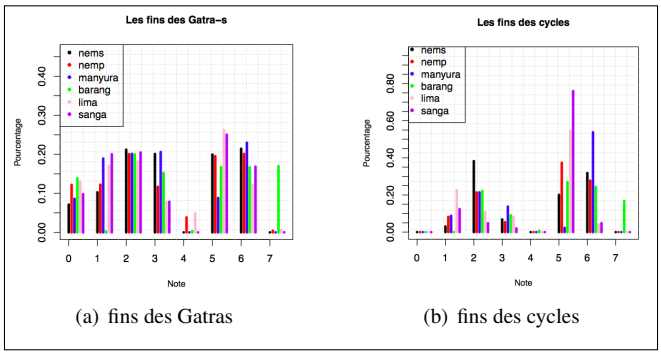
(a) nems (b) nemp (c) manyura

(d) barang (e) lima (f) sanga

FIG. 1 – *Pourcentage des notes dans les Gatras*

— les notes brutes $t_{k,i} \in \{n_j\}_{j=0..7}$
— les variations des notes successives $i+1$, $v_{k,i+1} = t_{k,i+1} - t_{k,i}$
— les fréquences des notes $f_{k,i} = \frac{\sharp t_{k,i}}{\sum \sharp n_j}$
— les probabilités des transitions entre notes $P(t_{k,i+1} = n_2 | t_{k,i} = n_1)$

Afin de mieux comprendre la répartition des notes dans les Gatras en fonction du Pathet, nous avons étudié tout particulièrement les histogrammes des fréquences (cf. Figure 1). Les résultats montrent clairement qu'un même histogramme peut être identique pour deux Pathets différents. Or deux Pathets peuvent être identiques en termes de fréquences des notes mais différents rythmiquement par rapport à l'intensité et la longueur des variations. Le Gatra est par conséquent un bon descripteur de mélodie puisqu'il évoque une notion topologique ou de forme mélodique et cependant ne permet pas d'inclure la dimension rythmique. Contrairement à la mélodie, le rythme est une notion temporelle marquée par les instants clés où les Gatras sont joués. Les résultats des histogrammes calculés spécifiquement en fin des Gatras et en fin des cycles par Pathet corroborent cette constatation (cf. Figure 2). Nous distinguons ainsi deux catégories de descripteurs : de mélodies et de rythmes. L'impact et la pertinence de ces descripteurs pris individuellement dans la classification des Pathets sont étudiés. L'utilisation mutuelle de ces descripteurs dans les mêmes modèles de classification fait l'objet d'une autre étude.



(a) fins des Gatras (b) fins des cycles

FIG. 2 – *Histogrammes des fréquences des notes en fin des Gatras et en fin des cycles par Pathet*

Descripteurs de mélodies

Les descripteurs associés à la mélodie sont tous les descripteurs orientés Gatra. Nous supposons que les Gatras d'une même pièce sont indépendants dans l'espace (le rang des notes) et dans le temps (les instants où les notes sont jouées). Les pièces sont vues comme une liste aléatoire de Gatras. Les observations sont alors des simples Gatras pour la classification des Pathets.

Descripteurs de rythmes

Dès lors que l'ordre des Gatras est important rythmiquement, des descripteurs orientés cycle sont choisis. Étant donnée la taille variable des cycles, nous avons choisi de prendre des blocs de 4 Gatras successifs puisque la taille d'un cycle est souvent multiple de 4. Sur ces cycles, nous considérons les descripteurs liés au : (1) fin des Gatras de fin des cycles ; (2) Gatra du milieu d'un cycle ; (3) Gatra du $\frac{1}{4}$ d'un cycle ; (4) Gatra du $\frac{3}{4}$ d'un cycle ; (5) fins de tous les Gatras d'un même cycle ; (6) la matrice des fréquences des notes ; (7) la matrice des transitions. Nous supposons que les blocs sont indépendants et constituent les observations pour la classification des Pathets.

2.3 Sélection de descripteurs par une approche comparative de modèles de classification

Nous menons dans cette étude une stratégie experte qui a pour objectif d'identifier les meilleurs descripteurs permettant d'obtenir les meilleurs modèles prédictifs de la classe d'une pièce en tenant compte des connaissances des experts. Pour ce faire, nous avons choisi de mettre en compétition trois algorithmes de classification : un modèle de régression linéaire AFD, un modèle bayésien MAP et un modèle support vecteur machines (SVM). Les trois algorithmes diffèrent par leur critère de classification. Les descripteurs qualifiés pertinents sont ceux qui permettent aux trois algorithmes de converger vers des erreurs de classification faibles en phase d'apprentissage et de test. L'évaluation de la performance d'un tel modèle de classification est une étape importante pour la compréhension de la qualité du mo-

dèle et aussi pour la sélection du modèle le plus approprié. Étant donnée la matrice de confusion d'un classifieur, la justesse d'une classification peut être évaluée par plusieurs mesures : le nombre d'exemples correctement classés positifs (TP) ; le nombre d'exemples correctement classés négatifs (TN), le nombre d'exemples incorrectement attribué à une classe positive (FP) et le nombre d'exemples positifs qui sont classés négatifs (FN). En plus de ces erreurs, nous calculons les 3 critères classiques d'évaluation de la performance : *le rappel*, *la précision* et la mesure *F* (F-Score). La précision mesure la proportion d'exemples bien classés parmi tous les exemples classés alors que le rappel (sensibilité) mesure la proportion d'exemples bien classés parmi tous les exemples de la base. Un classifieur parfait aura une précision et un rappel égaux à 1. En réalité, les algorithmes de classification sont plus ou moins précis, et plus ou moins justes. Il sera possible d'obtenir un système très précis ($Precision \approx 0,99$), mais peu performant ($rappel \approx 0.10$). Étant donnée une matrice de confusion, d'après (Sokolova et Lapalme, 2009)(Razik et al., 2003), le rappel et la précision sont définie, pour une classe C_i, par $Rappel(C_i) = \frac{TP_{C_i}}{TP_{C_i}+FN_{C_i}}$. La précision par classe est définie par : $Précision(C_i) = \frac{TP_{C_i}}{TP_{C_i}+FP_{C_i}}$. La qualité d'un classifieur est évaluée par les mesures moyennes : $Précision = n^{-1}\sum_{i=1}^{n} Précision(C_i)$; et le rappel moyen : $Rappel = n^{-1}\sum_{i=1}^{n} Rappel(C_i)$. $F1$-mesure ou F-score est une combinaison pondérée de la précision et le rappel : $F_1 = \frac{2\cdot(\text{Précision}\cdot\text{Rappel})}{(\text{Précision}+\text{Rappel})}$

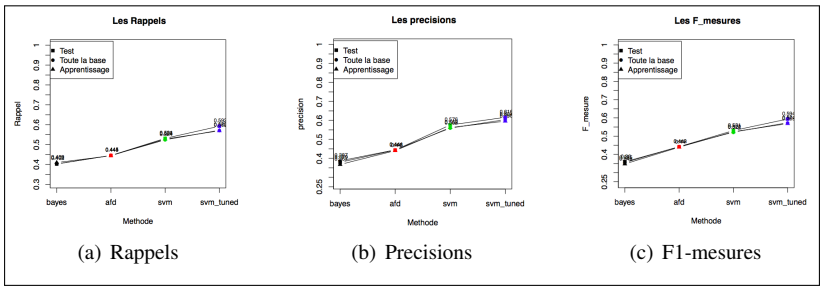
3 Validation expérimentale

3.1 Jeu de données

Dans le cadre de nos expérimentations, nous avons construit une base de données de type MangoDB. Cette base nous a permis de saisir 313 pièces différentes composées de 4592 blocs et 18712 Gatras (cf. tableau 2). Les pièces appartiennent à des Pathets différents divisant le jeu de données en 6 classes de Pathet et par échelle (Trois classes par échelle) de cardinalités comparables. Ce jeu de données constitue notre vérité terrain pour les tâches de reconnaissance et de classification des pièces. Deux échantillons aléatoires disjoints constitué chacun de la moitié du jeu de données ont servi respectivement à l'apprentissage des trois modèles et aux tests.

Classe	Nb.Pièces	Nb.Gatras	Nb.Blocs
$barang$	53	3383	835
$Manyura$	50	2803	684
$Lima$	59	2934	728
$Nemp$	50	3282	808
$Nems$	50	3276	799
$Sanga$	51	3034	738
$Total$	313	18712	4592

TAB. 2 – *Répartition des données*



(a) Rappels (b) Precisions (c) F1-mesures

FIG. 3 – *Récapitulatif des Performances des Gatras de type Pelog : Rappels, Précisions et les mesures de* F_1

3.2 Protocole expérimental

Chaque pièce est assignée à une classe de Pathet et une échelle. L'ensemble des pièces est divisé en deux échantillons de tailles égales d'apprentissage et de test. Afin d'éviter un quelconque biais dans nos expérimentations, les observations sont tirées aléatoirement et les performances des trois algorithmes d'apprentissage sont comparés sur la base du même échantillon et ce pour chaque descripteur. En d'autres termes, pour chaque descripteur sélectionné (décrits dans la section 2.2, page 4), les matrices de confusion, le rappel, la précision (eq. 2.3) et la F-mesure (eq. 2.3) sont calculés selon les trois modèles : AFD, le modèle Bayésien et SVM (Nous avons également regardé le modèle tuned SVM avec sélection de variables). Nous avons appliqué ce même procédé pour la reconnaissance du Pathet avec ou sans distinction de l'échelle. Quel que soit le descripteur, l'intégration du label échelle permet d'améliorer considérablement les performances des classifications.

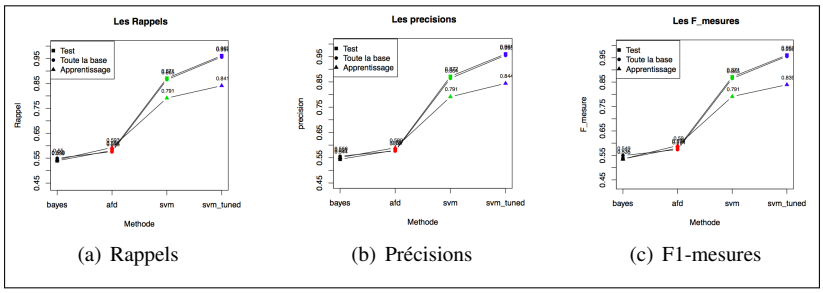
3.3 Résultats expérimentaux et validation par les experts

Performances de la classification avec les Gatras

En étudiant les Gatras sans connaissances *a priori* de l'échelle, nous remarquons que les performances en Rappel, en Précision et en F_1 restent entre 24% et 37%. Les Pathets de Pelogs sont séparables avec des mesures entre 40% et 60%. En Slendro, les mesures sont entre 39% et 54%. Les Gatras ne sont donc pas des descripteurs adéquats pour la reconnaissance des Pathets. Le deuxième choix intuitif, est d'utiliser les cycles pour décrire les partitions. Toutefois, les cycles du Gamelan sont de tailles différentes, nous avons donc choisi de diviser les cycles longs en Blocs de 4 Gatras (quatre Gatras est la longueur minimale d'un cycle). L'idée étant de supposer que les Gatras ne contiennent pas suffisamment d'information, peut-on trouver dans ce contexte suffisamment d'information dans les Blocs.

Performances de la classification avec les Blocs

Comme prévu, les Blocs apportent plus d'informations que les Gatras pour la reconnaissance des Pathets, bien que ces informations restent non satisfaisantes. Les performances des classifications des Blocs illustrent ceci avec une augmentation des rappels, des précisions et de la mesure F_1 égales à 35% par bayes et par AFD contre 24% avec les Gatras. Les machines à



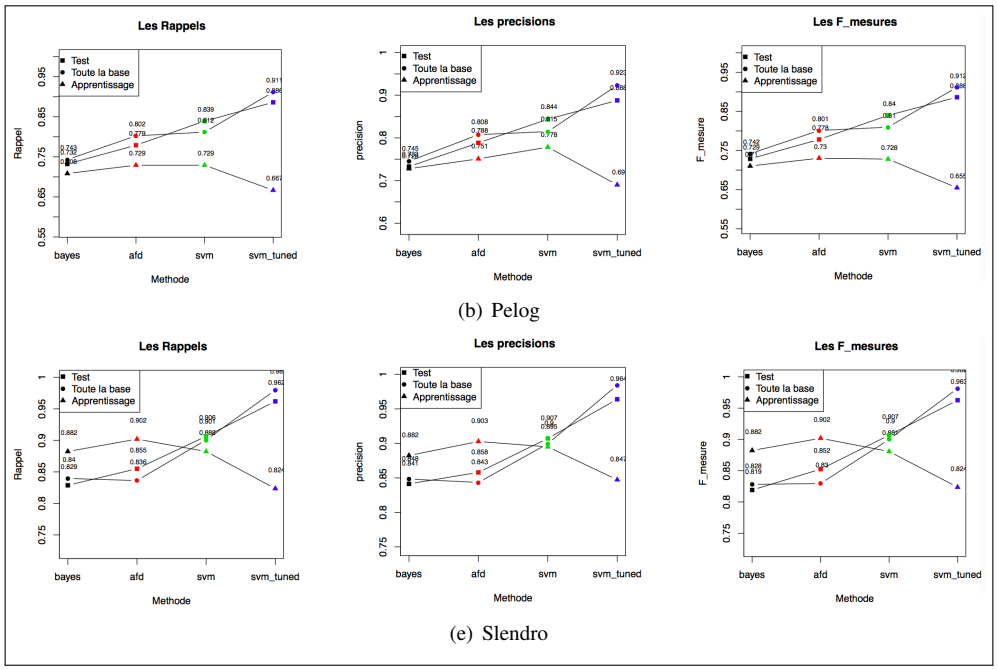
(a) Rappels (b) Précisions (c) F1-mesures

FIG. 4 – *Récapitulatif des mesures de Performances avec les blocs : Rappels, Précisions et les mesures de F_1 pour le plego*

vecteurs supports donnent une meilleure performance que les méthodes probabilistes avec 75% à 76% en apprentissage et 66% en test. Avec tune-SVM, les performances atteignent les 90% en apprentissage et les 80% en test. Une amélioration des performances est observée lorsque nous séparons les Blocs Pelog des Slendro. Les trois mesures varient entre 53% et 55% avec la classification Bayésienne. L'AFD, atteint les 59% pour les trois mesures. De la même manière, les SVM apprennent à reconnaître les Pathets Pelog jusqu'a 87% sans sélection de paramètres et 96% avec sélection de paramètres. En test, les SVM arrivent à prédire le Pathet jusqu'à 79% sans sélection de paramètres et 84% avec sélection de paramètres.

Performances de la classification avec les fréquences

L'amélioration des résultats de reconnaissance des pathets est très remarquable par rapport aux Blocs. En effet, la classification bayésienne arrive à reconnaître et à prédire les Pathets avec des mesures de performances stables en apprentissage et en test de l'ordre de 70%. L'AFD arrive à reconnaître et prédire les Pathets avec des mesures stables de 66% à 72% en apprentissage et de 61% à 69% en test. Nous avons également remarqué lors de l'interprétation de la projection en AFD des classes, que le Barang et le Lima se distinguent quasi parfaitement des autres Pathets. En Slendro, le Sanga et le Nem de Slendro se distinguent plus facilement par toutes les méthodes en apprentissage et en test alors que le Manyura est souvent confondu avec un Nem. Quant au SVM, ces mesures atteignent les 75% pour le rappel, 76% en précision et une mesure F_1 de 72% ; en test les trois mesures sont de l'ordre de 63%. D'après les résultats des classifications des Pathets selon l'échelle, on remarque que le Barang se distingue par toutes les méthodes en apprentissage et en test. La confusion reste, bien que légère, entre le Lima et le Nem de Pelog. En particulier, il y a moins de Nem qui sont confondus pour Lima que des Lima confondus pour des Nem. De ce fait, le Nem se distingue mais la confusion reste sur le Lima qui s'approche des Nem en termes de fréquences des notes. Les performances obtenues par les trois méthodes de classification pour les deux échelles sont similaires. En effet, la classification Bayésienne arrive à reconnaître et prédire les Pathets avec des mesures de performances stables en apprentissage de 84 à 88% et à 81% en test. Quant à L'AFD, elle arrive à reconnaître et prédire les Pathets avec des mesures stables de 83% à 90% en apprentissage et de 83% en test. SVM permet d'obtenir un rappel, une précision et une mesure F_1 de l'ordre de 90% en apprentissage et de 88% stable en test (cf. Figure 5).



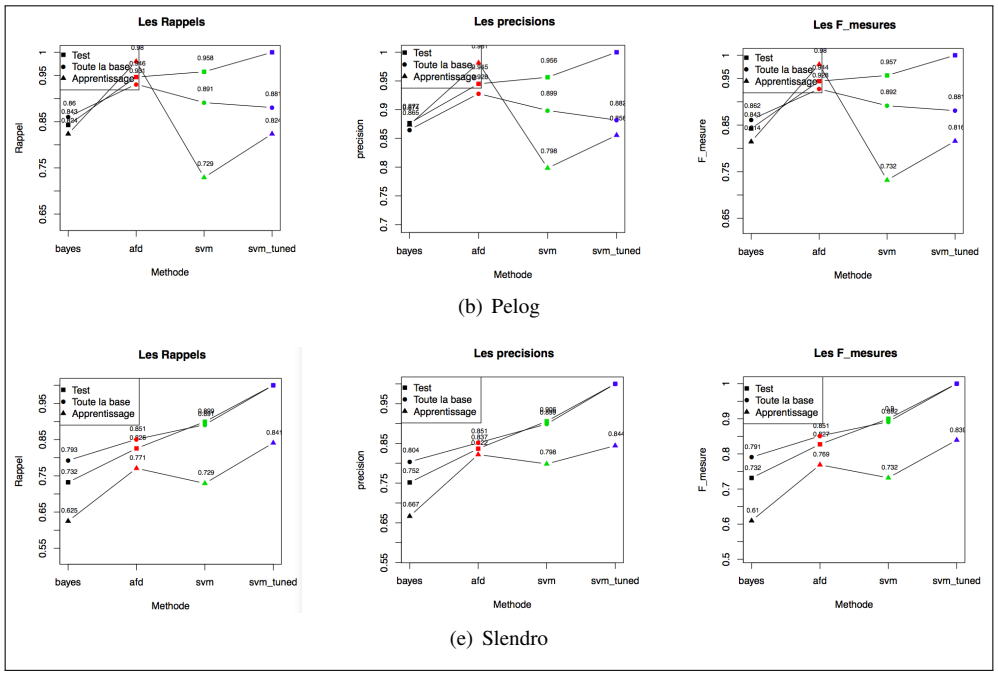
(b) Pelog

(e) Slendro

FIG. 5 – *Récapitulatif des mesures de Performances (b) Pelog Vs. (e) Slendro : Rappels, Précisions et les mesures de* F_1

Performances de la classification avec les matrices des transitions

La classification bayésienne arrive à reconnaître et à prédire les Pathets avec des mesures de performances stables en apprentissage et en test de 60 à 70%. L'AFD arrive à reconnaître et prédire les Pathets avec des mesures stables de 84% à 89% en apprentissage et de 76% à 71% en test. Avec SVM, le rappel, la précision et la mesure F_1 sont de 90% en apprentissage et de l'ordre de 72% en test. La projection en ACP des probabilités de transitions permet de distinguer le Barang par des probabilités élevées des occurrences de deux notes successives 21, 16, 10 et 47. La projection en AFD arrive à séparer parfaitement les partitions en Pelog des partitions en Slendro. Les projections sur le deuxième plan principal et le deuxième plan factoriel permettent de caractériser le Lima d'un côté avec l'ACP et de séparer le Lima et le Pelog des autres Pathets. Ces résultats se sont davantage améliorés avec la séparation des Pathets par échelle. En effet la classification Bayésienne arrive à reconnaître et prédire les pathets de type Pelog ou Slendro (cf. Figure 6) avec des mesures de performances stables entre 82 et 86% en apprentissage et de 84% en test. L'AFD arrive à reconnaître et prédire les Pathets avec des mesures stables de 92% à 98% en apprentissage et de 93% en test. Avec SVM, ces mesures ont un rappel, une précision et une mesure F_1 de 98% à 95% en apprentissage et de 72% à 79% en test.



(b) Pelog

(e) Slendro

FIG. 6 – *Récapitulatif des mesures de Performances des matrices de transitions en (b) Pelog et (e)Slendro : Rappels, Précisions et les mesures de* F_1

4 conclusion et perspectives

Les données du Gamelan Javanais sont particulièrement difficiles à caractériser et l'analyse fine des résultats qui nous avons proposé a permis de confirmer certaines conclusions attendues et d'atteindre de nouvelles qui étaient jusqu'à présent, considérées par les experts comme des hypothèses. Le protocole de test consiste à classifier par trois classificateurs différents les données décrites par les différents jeux de descripteurs. les descripteurs qualifiés pertinents sont ceux qui permettent aux algorithmes de classification de converger vers un modèle à moindre erreur (au sens des mesures classiques). Ces travaux ont ouvert la voie à diverses perspectives. Tout d'abord, de transcrire ces connaissances sous la forme de règles qui permettront de générer artificiellement des pièces et de les confronter à des experts du terrain. Ensuite de transposer cette étude directement sur des pièces sonores et de permettre la détection et la reconnaissance de motifs sonores clés. Enfin de proposer une approche par les chaînes de Markov cachées (HMM) afin de mieux modéliser l'aspect temporel et la dépenses entre blocs.

Références

El Khoury, E., S. Meignier, et C. Senac (2008). Segmentation et regroupement en locuteurs pour la parole conversationnelle. In *Journées d'Etudes sur la Parole (JEP), Avignon, France,*

09/06/2008-13/06/2008, http ://www.afcp-parole.org/, pp. 345–348. Association Francophone de la Communication Parlée (AFCP).

Fontan, L., C. Magnen, J. Tardieu, I. Ferrané, J. Pinquier, J. Farinas, P. Gaillard, et X. Aumont (2015). Comparaison de mesures perceptives et automatiques de l'intelligibilité : application à de la parole simulant la presbyacousie. *Traitement Automatique des Langues 55*(2), 151–174.

Guerid, A., A. Houacine, R. André-Obrecht, et H. Lachambre (2012). Performance of new voice conversion systems based on GMM models and applied to Arabic language. *International Journal of Speech Technology 15*(4), 477–485.

Razik, J., C. Senac, D. Fohr, O. Mella, et N. Parlangeau (2003). Comparison of two Speech/Music Segmentations Systems for Audio Indexing on the Web. In *7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003), Orlando, Florida, USA, 27/07/2003-30/07/2003*, ISBN : 980-6560-06-X. IIIS.

Sokolova, M. et G. Lapalme (2009). A systematic analysis of performance measures for classification tasks. pp. 427–437.

Souviraà-Labastie, N., A. Olivero, E. Vincent, et F. Bimbot (2015). Multi-channel audio source separation using multiple deformed references. *IEEE Transactions on Audio, Speech and Language Processing 23*(11), 1775–1787.

Summary

The Javanese Gamelan is essentially of metallophones orchestra. Each gamelan has its own agreement, which carcatérise his musical identity in relation to other gamelan in the same category as a community signature sound associated to the instrument. The music of gamalan is transmitted through learning and rouphly written. Indeed, there are simply music sheets memory aids to partially transcribe the music played. These are incomplete but are high information content.

Du fichier audio à l'intonation en Français : Graphes pour l'apprentissage de 3 classes intonatives

Martine Cadot*, Anne Bonneau**

* LORIA/UdL,
54600 Villers-lès-Nancy, France
martine.cadot@loria.fr,
http://www.loria.fr/ cadot
**LORIA/CNRS,
54600 Villers-lès-Nancy, France
anne.bonneau@loria.fr,
http://www.loria.fr/ bonneau

Résumé. L'analyse de l'intonation en Français est une tâche complexe. Partant de l'enregistrement audio d'une personne prononçant une phrase, il s'agit d'en extraire les paramètres prosodiques, puis de retrouver les différents types d'information que le locuteur a essayé de transmettre par ses changements d'intonation : phrase interrogative, déclarative, ou autre. Nous présentons ici en détail le processus complet de catégorisation statistique en trois types de phrases, de la collecte des données à l'interprétation des résultats, en passant par l'exposé de la méthode de traitement adaptée à notre problématique.

1 Introduction

Malgré le grand nombre d'études consacrées à l'intonation du français, relativement peu d'entre elles ont porté sur une analyse fine des indices acoustiques et sur un grand nombre de locuteurs (Vassière, 2002; Di Cristo, 1998). En outre, la reconnaissance automatique des différentes intonations est actuellement faite par des systèmes de reconnaissance fondés sur des modèles cachés, qui ne dévoilent pas les paramètres utilisés pour leurs résultats, et qui ont pour l'instant des résultats assez faibles. L'interprétation objective des indices prosodiques, et tout particulièrement des courbes mélodiques est actuellement de plus en plus utilisée pour l'apprentissage des langues étrangères et, dans le domaine de la rééducation, l'apprentissage de la langue maternelle, voir De Bot (1983) et Bonneau et Colotte (2011). Dans une première partie nous décrivons la problématique, dans la suivante les données et leur recodage, puis dans la troisième leur traitement, et enfin l'interprétation des résultats.

2 La problématique

Notre but est de construire un modèle statistique permettant de différencier les types intonatifs d'une phrase prononcée par un locuteur à partir des indices acoustiques de son enregis-

trement sonore. Le modèle que nous visons est de la forme X–>Y, se lisant selon les cas : X implique Y, X cause Y, X explique Y, X produit Y, etc. [1] où X représente les indices acoustiques et Y le type intonatif de la phrase. Il est "asymétrique" dans la mesure où les variables n'ont pas toutes le même rôle [2] : la variable Y est la "variable à expliquer" (appelée également variable dépendante), et les variables acoustiques X sont les "variables explicatives" (appelées également variables indépendantes).

Pour construire ce modèle nous avons utilisé une partie des données recueillies selon un "plan d'expérience" pré-établi [3], nous avons créé 29 variables acoustiques quantitatives X, 3 variables qualitatives (appelées également catégorielles) Z à contrôler (telles que le genre : M/F, dont il faut contrôler les effets sur le modèle pour les retirer le cas échéant [4]) et la variable catégorielle de contour mélodique Y dont nous n'avons gardé que 3 modalités.

Les données et leur structuration au sein d'un plan d'expérience seront décrites dans la section suivante. Nous définissons ci-dessous les variables X, Y et Z.

2.1 Le type intonatif : variable Y à expliquer

Lors d'une conversation en français entre deux personnes, il n'y a pas que les mots qui donnent du sens, il y a aussi l'intonation : par exemple si le ton du locuteur monte en fin de phrase, l'auditeur en déduit généralement qu'on lui pose une question. Delattre (1966) distingue quatre niveaux de hauteur de voix pour chaque locuteur, et construit à partir de là une typologie à 10 classes. Nous nous appuyons sur sa méthodologie pour définir trois types mélodiques. Voici leurs définitions, avec pour chacun un exemple relatif au syntagme "nos amis" et une figure dans laquelle la partie correspondant à "nos amis" est encadrée de marron.

1. les questions sans marques grammaticales, que nous notons ***qis***, avec un contour montant qui part du niveau 2 pour aller au niveau 4
   — Ex : la question "nos amis ?" à la fin de la phrase "Qui a téléphoné, ***nos amis ?***", voir la courbe bleue Fig.1.
2. les phrases déclaratives, que nous notons ***dis***, avec un contour (final) descendant, partant du niveau 2 pour aller au niveau 1
   — Ex : la réponse "Nos amis.", à la question "Qui arrive ?", voir la courbe bleue Fig.2.
3. les syntagmes non terminaux, que nous notons ***cis***, avec un contour montant partant du niveau 1 pour aller au niveau 2
   — Ex : "Nos amis" qui est au début de la phrase "***Nos amis*** aiment bien Nancy", voir la courbe bleue Fig.3.

2.2 Les indices acoustiques : variables X explicatives

Le principal est la hauteur de voix, notée F0 car c'est la "fréquence fondamentale", indiquée par des points bleus sur les figures1, 2 et 3. Nous disposons des valeurs successives de

1. Pour une sémantique de l'implication statistique, voir l'introduction de l'ouvrage de Gras et al. (2013).
2. Pour un exposé détaillé sur la planification des variables en psychologie expérimentale voir Hoc (1983), en agronomie voir Dagnelie (2003).
3. Pour plus de détails sur les plans d'éxpériences, voir Winer et al. (1991).
4. Pour un exposé détaillé sur la gestion des variables à contrôler en psychologie expérimentale voir Léon et al. (1977), en médecine voir Schwartz (1991).

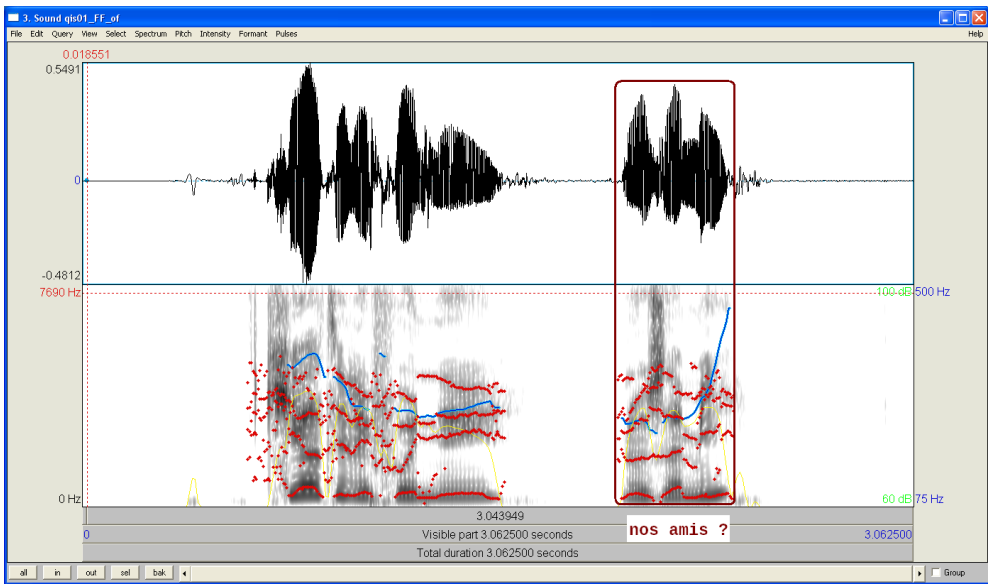


FIG. 1 – *Visualisation avec le logiciel PRAAT (2015) de la phrase "Qui a téléphoné,* **nos amis ?***", de type* **qis** *prononcée par la locutrice d'identifiant "of". Points bleus : F0, jaunes : Intensité, rouges : Formants.*

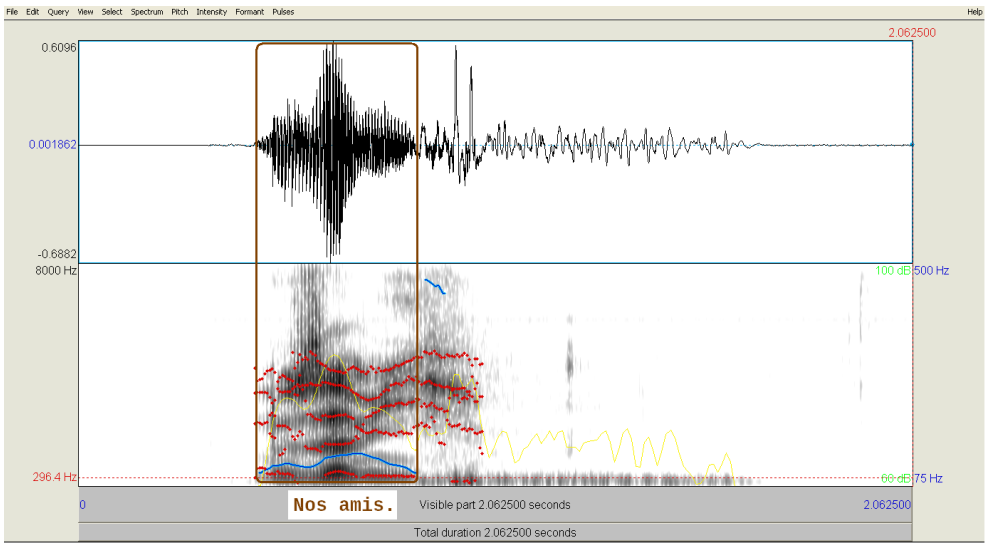


FIG. 2 – *Visualisation avec PRAAT de la phrase "* **Nos amis.***", de type* **dis** *prononcée par le locuteur d'identifiant "vv". Points bleus : F0, jaunes : Intensité, rouges : Formants.*

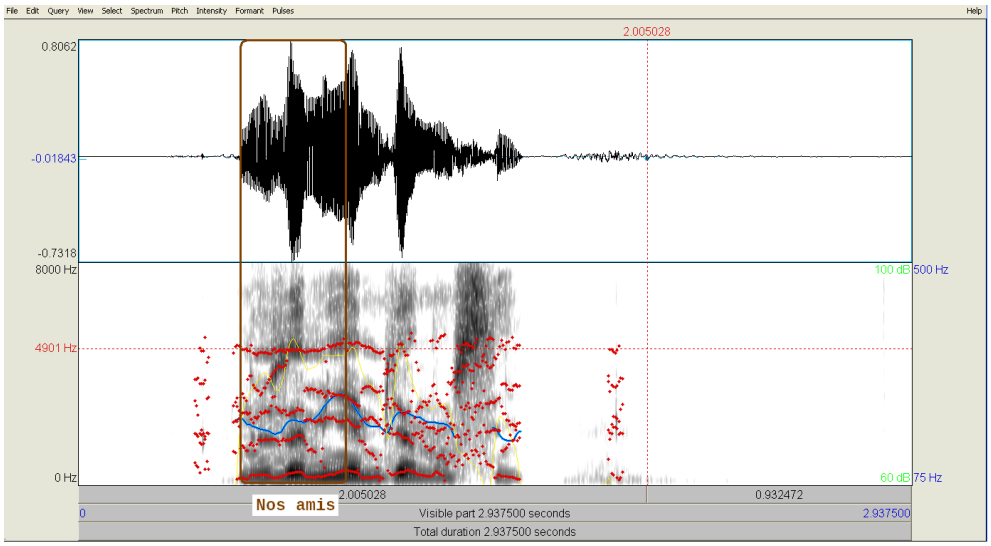


FIG. 3 – *Visualisation avec PRAAT de la phrase "***Nos amis** *aiment bien Nancy", de type* **cis** *prononcée par le locuteur d'identifiant "tl". Points bleus : F0, jaunes : Intensité, rouges : Formants.*

F0 pour chaque phrase prononcée, leur nombre variant selon le contexte et le locuteur. Chaque phrase a été segmentée de façon automatique [5].

Nous avons calculé 9 indices acoustiques par son à partir de scripts PRAAT que nous avons écrits. Nous les avons obtenus en appliquant diverses fonctions comme **max**, **min**, etc. à l'ensemble des valeurs successives de F0 pour un même son. Certains de ces indices ont une distribution de valeurs très déséquilibrée, avec parfois quelques valeurs extrêmes. Celle qui a le plus grand nombre de valeurs extrèmes, qu'elles soient positives ou négatives, est la pente d'un son. Ce n'est pas surprenant, son calcul s'appuyant sur la dérivée de F0 par rapport au temps, donc délicate à estimer quand il y a des points de discontinuité.

Pour cette étude, nous nous sommes limitées à l'étude des sons correspondant aux trois voyelles des syntagmes étudiés, numérotées par leur rang dans la chaîne sonore, numéro 3 pour la première (le 'o' de "nos amis"), 5 pour la deuxième (le 'a' de "nos amis") et 7 pour la dernière (le 'i' de "nos amis"). Nous avons ainsi $9 \times 3 = 27$ indices acoustiques auxquels s'ajoutent deux indices obtenus par différence de valeur des précédents sur 2 voyelles voisines (3 et 5, 5 et 7), ce qui nous donne les 29 variables suivantes : 3maxF0, 5maxF0, 7maxF0, 3minF0, 5minF0, 7minF0, 3meanF0, 5meanF0, 7meanF0,..., maxF0_v5_3, maxF0_v7_5.

5. Le logiciel PRAAT (2015) et le logiciel ProsodyPro (2015) de Yi Xu ont été d'une grande aide pour la segmentation en sons et la récupération des indices acoustiques. Toutefois, comme on peut l'imaginer en regardant les figures 1, 2 et 3, un certain nombre de vérifications et de corrections on été nécessaires avant d'envisager le traitement des données numériques.

2.3 Les variables Z à contrôler

Elles sont essentiellement de 2 groupes. Le premier groupe concerne le locuteur : homme ou femme, âge, etc.

Le second groupe de variables à contrôler concerne les différentes phrases à prononcer. Les indices acoustiques X étant calculés sur les voyelles de rangs 3, 5 et 7, il convient de contrôler leurs variations quand on remplace pour un rang donné une voyelle par une autre. Et les consonnes qui sont aux rangs pairs doivent également être contrôlées parce que l'intonation est portée par certaines consonnes plus que par d'autres [6], ce qui produit ou non un "débordement" du F0 sur la voyelle suivante .

A ces deux groupes spécifiques de variables s'ajoute le groupe des variables à contrôler habituelles dans ce type d'expérience : celui des variables "d'environnement" liées à la passation de l'expérience, telles que changement de matériel, de conditions, etc. lors de l'enregistrement.

3 Les données et leur structure

Ce corpus est issu d'un corpus plus vaste, qui comprend notamment des phrases plus longues, et qui a été enregistré dans le cadre du projet Intonale (2009-2011) supporté par le Comite de Coordination et d'Orientation Scientifique Lorrain : Nancy University, Paul Verlaine University, CNRS,INRIA, INRA, INSERM, CHU.

Nous ne décrivons ici que les éléments nécessaires à notre étude. Celle-ci s'appuie sur 115 enregistrements sonores concernant 5 phrases simples prononcées de 3 façons différentes par 33 locuteurs différents, chaque locuteur n'ayant prononcé qu'une partie des phrases. Nous décrivons d'abord les conditions de l'expérience construite pour enregistrer la parole des locuteurs, puis son effet sur la structuration des données.

3.1 Les conditions de l'expérience

Les éléments relatifs aux 115 enregistrements que nous avons analysés sont décrits en utilisant la terminologie des plans d'expérience (Hoc 1983).

Facteur P : Phrase. Les cinq phrases (ou parties de phrases) qui ont été étudiées sont :
P1 : "Nos amis", P2 : "Les élèves", P3 : "Les lamas", P4 : "La marelle" P5 : "Les agneaux"

Facteur T : Type de phrase. Le contexte dans lequel les phrases ont été lues a été conçu pour que les locuteurs prononcent les phrases de manière neutre, sans émotion particulière, avec l'intonation qui respecte chacune des trois catégories qui nous intéressent ici. Certains locuteurs ont cependant parfois donné à certaines phrases une émotion particulière, notamment en fin de phrases déclaratives (expression du doute, par exemple). Nous avons conservé ces phrases, et notre analyse a donc porté, sur les questions (qis), les syntagmes non terminaux (cis), et les déclarations (dis), ces dernières comportant un certain nombre de phrases non neutres. C'est ce facteur "type de phrase" qui deviendra la variable Y de notre modèle X–>Y

6. En effet, les consonnes "vocaliques" (le son 'r' de Laure) ont une fréquence fondamentale non nulle, contrairement aux consonnes "sourdes" (le son 'k' de Jacques).

Facteur L : Locuteur. Ce sont 6 hommes et 27 femmes, ayant le français comme langue maternelle. Le nombre de phrases que chaque locuteur a prononcées varie entre 2 et 5.

Facteur G : Genre. Ce facteur à 2 modalités H (Homme) et F (Femme). Son effet a été supprimé avant modélisation en retranchant aux valeurs de F0 leur moyenne par locuteur.

Facteur R : Rang du son. Pour rendre compte de la totalité du contour mélodique de chaque phrase, nous avons numéroté chaque partie de la phrase. Le numéro 1 correspond à ce qui est enregistré avant le début, le numéro 2 à la première consonne, le numéro 3 à la première voyelle et ainsi de suite, le numéro 7 correspondant à la troisième voyelle, et les numéros à partir de 8 à d'éventuels sons supplémentaires, ces derniers pouvant contenir une information pertinente. Par exemple, pour la phrase interrogative "Nos amis ?" les numéros vont de 1 à 7, par contre pour les syntagmes non terminaux (de type "cis"), les numéros vont au-delà de 7 car pour forcer le locuteur à adopter l'intonation convenable pour "nos amis", il lui a été demandé de prononcer la phrase plus longue "nos amis aiment bien Nancy" (voir Fig. 3).

Facteur S : Son. Son nombre de modalités dépasse 10 car il en a autant que de sons différents dans les 5 phrases. De plus on a vu qu'un même son pouvait se prononcer différemment selon le son précédent et le suivant, et c'est peut-être le facteur "2S", des suites de 2 sons (consonne+voyelle) qu'il faudrait considérer, avec un nombre de modalités encore plus important.

3.2 Le plan d'expérience

Nous aurions aimé que ce soit un plan factoriel complet équilibré, qu'il y ait autant d'hommes que de femmes, que chacun prononce les 5 phrases dans les 3 intonations, ou bien que ce soit un plan randomisé dans les règles de l'art afin d'éliminer aisément les effets des variables Z à contrôler sur la relation entre X et Y.

Ce n'a pas été le cas, ce qui s'explique par le poids de réalisation d'une telle expérience, autant pour ceux qui la subissent en tant que locuteur, que pour ceux qui la font passer et ceux qui transforment les enregistrements sonores en fichiers de données exploitables statistiquement. En effet, dans le cadre du projet Intonale, chaque locuteur a prononcé une dizaine de phrases, dont nous n'avons extrait qu'une petite partie, les plus simples, pour initier notre analyse. Puis c'est de façon semi-automatique que nous avons créé le fichier destiné aux analyses statistiques, avec 115 lignes et une cinquantaine de colonnes : il nous a fallu annoter les 115 enregistrements sonores, en extraire les parties correspondant aux phrases visées, et les aligner (leur longueur dépendant de la rapidité d'élocution du locuteur), avant de créer les variables.

Malgré cela, c'est une chance d'avoir pu obtenir de telles données, qui se sont avérées adaptées à notre problématique une fois choisis les indices acoustiques et la technique statistique appropriés. Pour notre étude, tous les facteurs de ce plan d'expérience ont été pris en compte dans le modèle X–>Y que nous cherchons à établir : **Rang du son** dans la création des variables formant X, **Phrase**, **Son**, **Locuteur** et **Genre** ont formé la matrice Z des variables à contrôler et **Type de phrase** est devenu la variable Y du modèle X–>Y à établir.

4 Le traitement des données

Nous cherchions une méthode statistique permettant de prédire Y, le type de phrase, variable qualitative à 3 modalités, à partir de X, matrice de 29 variables acoustiques, tout en contrôlant Z, matrice de variables qualitatives. Nous avons exploré les modèles possibles en statistiques classiques et en apprentissage automatique (Mitchell, 1997) avant de nous tourner vers l'ASI (Gras et al., 2013), plus susceptible de répondre à nos attentes.

Modèle X->Y en statistiques classiques Le "modèle linéaire" (Prum, 1997) des statistiques classiques [7], comparaisons de moyennes de l'ANOVA (Winer et al., 1991), régression linéaire multiple (Baillargeon, 2000), est particulièrement adapté aux données structurées selon un plan d'expériences, à condition toutefois que la variable Y soit quantitative continue. Notre variable à expliquer Y étant catégorielle à 3 modalités, le modèle linéaire ne pouvait s'appliquer, et les deux autres modèles des statistiques classiques compatibles avec cette contrainte que nous avons explorés, la régression logistique et l'analyse discriminante (Nakache et Confais, 2003) ne pouvaient s'utiliser avec les variables X explicatives que nous avons (trop nombreuses, distribution déséquilibrée de leurs valeurs), et les variables qualitatives Z à contrôler.

Modèle X->Y en apprentissage automatique Nous nous sommes alors tournées vers des méthodes de discrimination entre catégories fonctionnant par apprentissage automatique [8], moins exigeantes par nature sur le type de distribution des variables X. Nous avons d'abord sélectionné les "MSVM" (Weston et Watkins, 1998), qui sont une extension des Support Vector Machine à la prédiction d'une variable Y multi-catégorielle et non simplement binaire. Mais le caractère "boîte noire" des MSVM nous a conduites à rechercher des méthodes fournissant un modèle plus explicite. La méthode par "arbres de décision" (Breiman et al., 1984) semblait fournir un tel modèle : elle produisait 3 règles du type A->B, pour lesquelles B prenait une des 3 valeurs "cis", "dis", "qis", et A était formé d'intervalles de valeurs de quelques variables de X. Ce modèle ne nous a pas convenu car trop simpliste à nos yeux. Nous voulions connaitre le rôle de chaque variable de X dans la prédiction de Y, ainsi que leurs interactions.

Modèle en Analyse Statistique Implicative Nos essais précédents nous ont permis de redéfinir nos attentes : nous désirions un modèle contenant toutes les liaisons significatives entre les variables X, Y et Z, plutôt qu'un modèle de prédiction de Y par X, avec Z contrôlé. C'est finalement vers une méthode d'extraction des "implications" entre variables à partir des données, l'ASI (***l'Analyse Statistique Implicative***), que nous nous sommes tournées.

L'Analyse Statistique Implicative, méthode statistique créée par Régis Gras (Gras et collaborateurs, 1996) nous a permis de faire émerger un modèle correspondant à nos attentes, et d'interprétation aisée de par sa forme graphique (voir figure 4). Nous décrivons d'abord les principes de cette méthode, puis l'utilisation que nous en avons fait pour nos données et les résultats que nous avons obtenus.

7. Nous appelons ici "statistiques classiques" les méthodes permettant d'inférer à une population les observations faites sur un échantillon respectant certaines conditions.

8. Nous appelons ici "apprentissage automatique" les méthodes permettant d'établir un modèle des données en l'extrayant, le contrôlant et le validant sur des parties complémentaires des données.

4.1 Les principes de l'Analyse Statistique Implicative

Conçue au départ pour traiter des données de type questionnaire avec une dizaine de variables binaires et une trentaine d'individus, la méthode donnait une liste d'implications statistiquement significatives entre les variables. Elle était implémentée dans le logiciel CHIC qui produisait ces implications sous forme d'un graphique clair. Ses utilisateurs étaient essentiellement des enseignants de mathématiques en collège et lycée essayant de comprendre comment s'organisaient les apprentissages de leurs élèves, quelles notions devaient être assimilées pour permettre l'apprentissage d'autres notions. Depuis la méthode s'est enrichie, elle permet de traiter des données plus importantes, plus variées, et ses utilisateurs se sont multipliés, élargissant son utilisation à des communautés scientifiques variées. En plus de la version Windows du logiciel CHIC (2012), une version libre devrait bientôt être disponible parmi les "packages" du langage R [9].

Cette méthode se situe à la fois en fouille de données et en statistique classique. Elle permet de produire des "règles d'implication" ressemblant aux "règles d'association" de la fouille de données [10]. Ces règles sont de la forme A->B, où A et B sont deux "items" [11]. Comme pour les règles d'association ce ne sont pas toutes les liaisons entre variables qui sont extraites, mais seulement les plus remarquables. Pour l'ASI, ce sont les règles "significatives", dans le sens usuel en statistiques classiques, c'est-à-dire s'appuyant sur un modèle probabiliste : "l'implication statistique X->Y" devient significative dès qu'il y a "relativement" très peu de contre-exemples, c'est-à-dire de cas qui vérifient X sans vérifier Y. Le mot "relativement" renvoie à des hypothèses relatives à la distribution des effectifs théoriques, binomiale par défaut, ou de Poisson.

L'avantage de cette méthode par rapport l'extraction des règles d'association est que le nombre de paramètres à choisir est restreint (bien inférieur aux quelque cinquante indices de qualité des règles d'association) et que l'ensemble des règles extraites est représenté par un graphe avec relativement peu de flèches. Ces flèches représentent des relations de type "causal" entre les variables, ce qui permet un raisonnement abstrait sur les données, s'appuyant sur des successions de flèches entre les variables. Pour arriver à un graphe aussi maniable par l'utilisateur il a fallu faire des choix, notamment n'accepter qu'une variable en partie gauche de la règle, supprimer les raccourcis dus à la transitivité, éviter les cycles, etc.

Nous renvoyons le lecteur intéressé par plus de détails sur les fondements et les évolutions de cette méthode à Gras et collaborateurs (1996), Gras et al. (2013) ou aux actes des conférences ASI [12].

4.2 Application de l'ASI aux données

Nous avons transformé la variable "type de phrase" en 3 variables binaires "qis", "cis" et "dis" et fait de même pour chacune des variables Z à contrôler. Puis nous avons découpé chacune des 29 variables acoustiques en 3 variables binaires, la première pour les plus faibles valeurs correspondant à un tiers des effectifs, la deuxième pour les valeurs médianes correspondant au deuxième tiers des effectifs, et la dernière pour les plus fortes valeurs avec le dernier

9. Pour la description du logiciel "rchic" et les instructions d'utilisation, voir Couturier et Dhouibi (2015).
10. Pour un exposé plus détaillé des règles d'association, voir Cadot (2006).
11. Un item est une variable binaire dont la valeur 1 indique une propriété possédée par le sujet.
12. Voir le site de l'ASI : http://sites.univ-lyon2.fr/asi8/?page=0.

tiers des effectifs. Le nom de ces variables a été créé en ajoutant un suffixe 1, 2 ou 3 au nom de la variable de départ. Par exemple “3meanF0”, moyenne des valeurs consécutives de F0 pour un même son, a produit “3meanF01”, variable binaire correspondant au valeurs les plus faibles, “3meanF02”, pour les valeurs médianes et “3meanF03” pour les valeurs les plus fortes.

Les variables explicatives X, à expliquer Y et à contrôler Z ont ainsi produit une centaine de variables binaires. C’est sur cette matrice formée d’une centaine de lignes et de colonnes que nous avons appliqué CHIC en demandant la production d’un graphe d’implication. Nous avons obtenu la figure 4.

4.3 Analyse du graphe implicatif obtenu

Moins de la moitié des variables apparaissent dans la figure 4, et elles se présentent sous forme de 8 graphes d’implication indépendants, que nous avons numérotés de 1 à 8 en partant de la gauche.

Un graphe par modalité de Y. Les 3 modalités de Y se trouvent dans 3 graphes séparés avec quasiment les mêmes variables acoustiques, mais à des niveaux différents :

- Le graphe 1 contient la modalité “cis” et les valeurs **médianes** (suffixe 2) des niveaux moyen (mean), max et final de F0 pour le dernier son, ainsi que la valeur médiane de la différence entre le max de F0 sur la dernière voyelle et son max sur la voyelle précédente,
- Le graphe 2 contient la modalité “dis” et les valeurs **faibles** (suffixe 1) des mêmes variables ainsi que du niveau min de F0 pour le dernier son,
- Le graphe 3 contient la modalité “qis” et les valeurs **fortes** (suffixe 3) des mêmes variables que le graphe 1.

On retrouve bien les contours mélodiques attendus des 3 types de phrases qui se différencient par leur niveau de F0 sur la dernière voyelle, ainsi que par l’augmentation de F0 entre l’avant-dernière et la dernière voyelle, fort pour les “qis”, moyen pour les “cis” et faible pour les “dis”. Et le fait que ni la première voyelle ni la seconde et avant-dernière n’intervienne dans la détermination des types mélodiques.

Un graphe par niveau de la première voyelle et de la deuxième. Les autres graphes montrent les co-variations des indices acoustiques :

- Le graphe 4 associe les niveaux forts de la première voyelle pour F0
- Le graphe 5 associe les niveaux forts de la deuxième et avant-dernière voyelle pour F0
- Le graphe 6 associe les niveaux moyens de la première voyelle pour F0
- Le graphe 7 associe les niveaux moyens de la deuxième et avant-dernière voyelle pour F0
- Le graphe 8 associe les niveaux faibles des deux premières voyelles

5 Bilan

Disposant d’enregistrements sonores de locuteurs en train de prononcer des phrases de 3 types mélodiques, notés “cis”, “dis” et “qis”, nous avons cherché une méthode permettant de prédire le type de phrase à partir des données acoustiques.

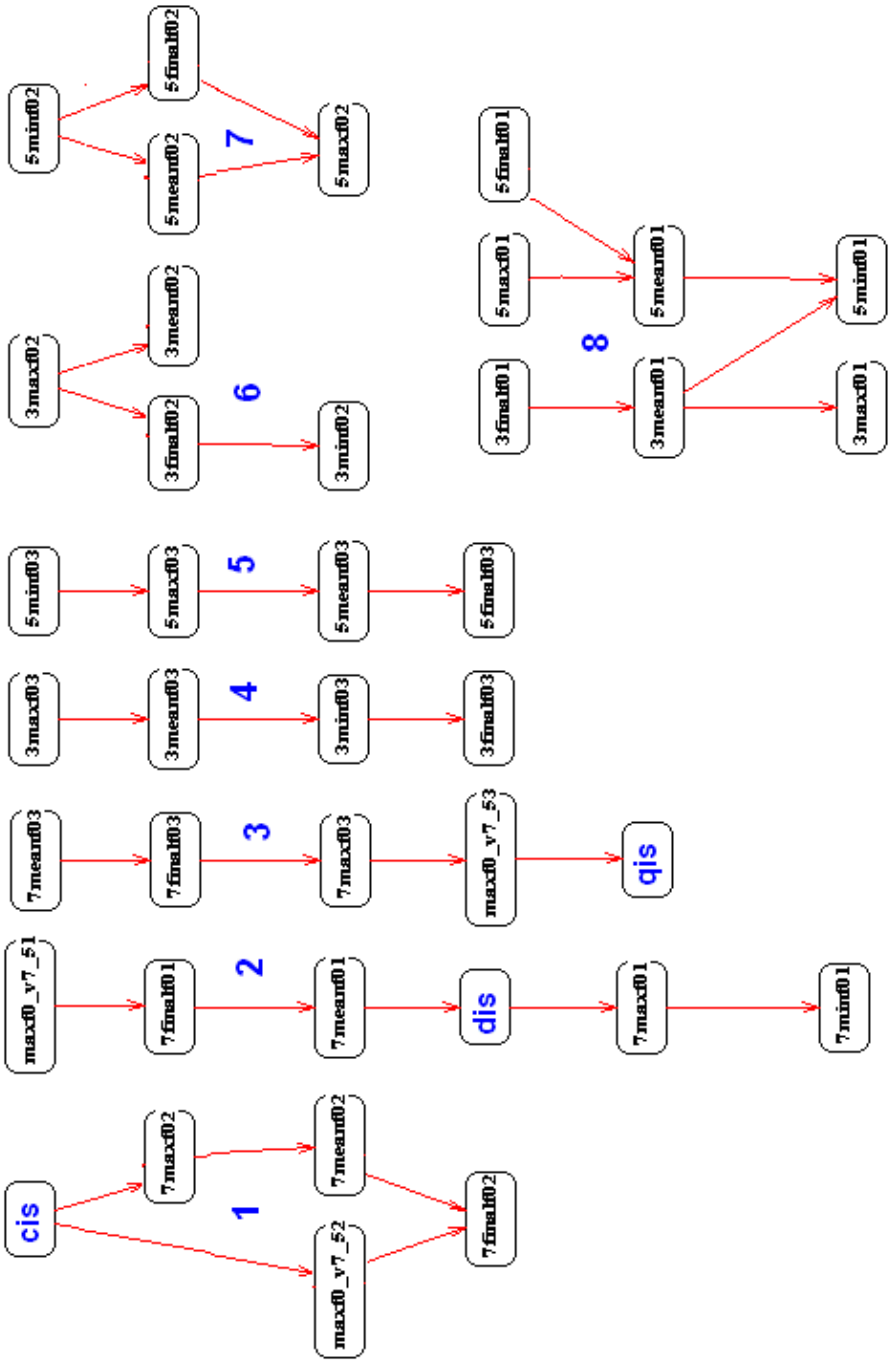


FIG. 4 – *Graphe des liens significatifs d'implication statistique extraits par CHIC sur nos données*

Les indices acoustiques ont été créés de façon semi-automatique à partir de F0 pour fournir un ensemble de valeurs exprimant au mieux l'évolution de l'intonation de chacun des sons successifs d'une phrase. Dans un but de simplification, nous nous sommes focalisées sur les 3 voyelles de chaque "phrase". La construction des indices acoustiques a abouti sur des variables aux distributions déséquilibrées, et très liées entre elles, ce qui nous a conduit à utiliser "l'Analyse Statistique Implicative" (Gras et collaborateurs, 1996) pour extraire un modèle explicite et opérationnel des données. Pour utiliser cette méthode, nous avons recodé les variables acoustiques en les transformant en variables qualitatives. Le choix d'une découpe en 3 parties de chaque variable [13] a permis d'obtenir des graphes séparés pour chaque type de phrase. Les flèches dans les graphes n'ont pas été interprétées de façon causale, mais en tant associations très significatives. Ainsi utilisé, le graphe implicatif nous a fourni les liens que nous attendions entre les variables acoustiques et les 3 courbes mélodiques.

Les 3 types de phrases choisis ici sont les plus courants. Il nous reste maintenant à étendre cette méthodologie à d'autres courbes mélodiques.

Références

Baillargeon (2000). *La régression linéaire, Edition des trois Sources*. Quebec : Edition des trois Sources.

Bonneau, A. et V. Colotte (2011). Automatic Feedback for L2 Prosody Learning. In I. Ipsic (Ed.), *Speech and Language Technologies*, pp. 55–70. Intech.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Cadot, M. (2006). *Extraction des liaisons complexes entre variables en Sciences Humaines : Statistiques, Motifs and Règles de'Association*. Theses, Université de Franche-Comté. Date de début de la rédaction : 2004.

Couturier, R. et A. Dhouibi (2015). Formation à l'A.S.I. par l'usage du logiciel CHIC et R-CHIC : retour sur quelques fonctionnalités. In *8ème Colloque International sur Analyse Statistique Implicative*, Radès, Tunisie, pp. 500–511.

Dagnelie, P. (2003). *Principes d'expérimentation : planification des expériences et analyse de leurs résultats.* Grenoble : Presses Agronomiques de Gembloux.

De Bot, K. (1983). Visual feedback on intonation i: Effectiveness and induced practice behaviour. *Language and Speech 6*(4), 331–350.

Delattre, P. (1966). Les dix intonations de base du français. *The french Review 40*, 1–14.

Di Cristo, A. (1998). Intonation in French. In D. Hirst et A. D. Cristo (Eds.), *Intonation systems : A survey of twenty languages*, pp. 195–218. Cambridge University Press.

Gras, R. et collaborateurs (1996). *L'implication statistique, une nouvelle méthode exploratoire de données*. Grenoble : La pensée sauvage.

Gras, R., J.-C. Régnier, C. Marinica, et F. Guillet (2013). *L'analyse statistique implicative : méthode exploratoire et confirmatoire à la recherche de causalités*. Toulouse : Cépaduès.

13. Nous avons testé un découpage en 4 parties, pour retrouver les 4 niveaux de Delattre, mais il s'est avéré improductif.

Hoc, J.-M. (1983). *L'analyse planifiée des données en psychologie*. Paris : Presses Universitaires de France.

Léon, A., J. Cambon, M. Lumbroso, et F. Winnykamen (1977). *Manuel de psychopédagogie expérimentale*. Paris : Presses Universitaires de France.

logiciel CHIC (2012). Classification Hiérarchique Implicative et Cohésitive. Version 6.0, copyright (c), http://www.ardm.eu/contenu/logiciel-d-analyse-de-donnees-chic.

logiciel PRAAT (2015). Praat: doing phonetics by computer. Version 6.0, copyright (c), http://www.fon.hum.uva.nl/praat/.

logiciel ProsodyPro (2015). A praat script for large-scale systematic prosody analysis. Version 5.5.7, http://www.homepages.ucl.ac.uk/ uclyyix/ProsodyPro/.

Mitchell, D. (1997). *Machine Learning*. New York: McGraw Hill.

Nakache, J.-P. et J. Confais (2003). *Statistique explicative appliquée : analyse discriminante, modele logistique, segmentation*. Paris : Editions Technip.

Prum, B. (1997). *Modèle linéaire. Comparaison de groupes et régression*. Paris : Eyrolles, Collection INSERM.

Schwartz, D. (1991). *Méthodes statistiques à l'usage des médecins et des biologistes*. Paris : Flammarion.

Vassière, J. (2002). Cross-linguistic prosodic transcription : French vs. English. In N. B. Volskaya, N. D. Svetozarova, et P. A. Skrelin (Eds.), *Problems and methods of experimental phonetics. In honour of the 70th anniversary of Pr. L.V. Bondarko*, pp. 147–164. St Petersburg State University Press.

Weston, J. et C. Watkins (1998). Multi-class support vector machine. Technica l report csd-tr-98-04, Royal Holloway, University of London, Department of Computer Science.

Winer, B., D. Brown, et K. Michels (1991). *Statistical principles in experimental design, 3rd edition*. New York: McGraw-Hill.

Summary

The analysis of the intonation of a sentence is a complex task. This implies the extraction of prosodic parameters from the recordings of sentences, then the analysis of the different kinds of information the speaker wants to give: statement, interrogative sentence, focus accent ...We present here in detail the process of the statistical categorization for three types of information conveyed by sentences (statement, yes-no question, and end of accentual groups), from the presentation of the corpus, then the method elaborated for our problem up to the interpretation of results.

Caractérisation d'instances d'apprentissage pour un méta-mining évolutionnaire

William Raynaut*, Chantal Soule-Dupuy*, Nathalie Valles-Parlangeau*,
Cedric Dray**, Philippe Valet**

*IRIT UMR 5505, UT1, UT3
Universite de Toulouse, France
prenom.nom@irit.fr
**INSERM, U1048
Universite de Toulouse, France
prenom.nom@inserm.fr

Résumé. Les nombreuses techniques d'apprentissage et de fouille de données peuvent se révéler d'importants atouts dans divers domaines, mais choisir la technique la plus appropriée pour une application précise est une tâche très complexe pour un non-expert. Notre objectif est ainsi de produire un assistant de modélisation répondant à ce besoin, par une approche à la frontière du méta-apprentissage et des heuristiques évolutionnaires. Nous présentons ici le fonctionnement prévu de cet assistant, suivi d'une discussion de notre approche du problème de caractérisation des instances d'apprentissage, qui reste un verrou majeur du méta-apprentissage et méta-mining.

1 Motivation

L'apprentissage a été un secteur très prolifique ces dernières décennies, produisant nombre de techniques et algorithmes. Cependant, leurs performances sont sujettes à d'importantes variations d'un jeu de données à l'autre. On retrouve ainsi dans les *"No free lunch theorems"* (Wolpert, 1996) l'idée qu'il n'existe pas de solution meilleure en toute situation, d'apprentissage meilleur dans tous les domaines. Firent suite nombre d'études catégorisant l'adéquation du biais d'algorithmes particuliers avec certains domaines ou situations, tels (Aha, 1992) et (Gama et Brazdil, 1995) qui y ont appliqué des techniques d'apprentissage de règles afin de décrire les conditions engendrant des différences de performance significatives entre algorithmes. Ces applications de l'apprentissage à l'étude de sa propre applicabilité ont posé les fondations du domaine du *méta-apprentissage*. Malgré bien d'autres applications fructueuses (Kalousis et Hilario, 2001b), et perspectives de recherches, telles la récursion infinie d'apprentissages auto-adaptatifs (Vilalta et Drissi, 2002), le problème du méta-apprentissage est toujours d'actualité, et les perspectives applicatives restent nombreuses. Parmi ces dernières, on considérera en particulier les récentes approches "*meta-mining*", qui s'affranchissent du cadre restreint de l'apprentissage au meta-niveau pour adresser une question plus large : *Considérant un besoin et un jeu de données, quelle suite de traitements donnera les meilleurs résultats ?*

Un algorithme est considéré capable d'apprentissage si sa performance s'améliore avec l'expérience (Mitchell, 1997). Dans le cas du méta-apprentissage, la plupart des approches impliquent la génération de méta-données, c'est-à-dire de données sur le problème de l'apprentissage, qui sont ensuite utilisées pour entrainer un algorithme d'apprentissage au méta-niveau. Notre perspective sur le sujet est que l'on peut définir un ensemble de méta-données comme une population de méta-instances d'apprentissage, chacune décrivant l'application de traitements précis à un jeu de données particulier, et qu'une bonne solution à un problème de méta-apprentissage peut être obtenue par l'emploi d'une heuristique évolutionnaire au sein de cette population. Le potentiel d'une telle approche n'a, à notre connaissance, pas encore été étudié, mais des approches comparables présentent des résultats encourageants, notamment pour des problèmes de satisfiabilité (SAT) (Xu et al., 2012) ou de sélection d'instance (Leyva et al., 2015).

Nous envisageons donc l'application de cette méthodologie à un assistant de modélisation, dont le fonctionnement prévu est illustré par la figure 1 et est présenté de manière plus complète dans (Raynaut et al., 2015).

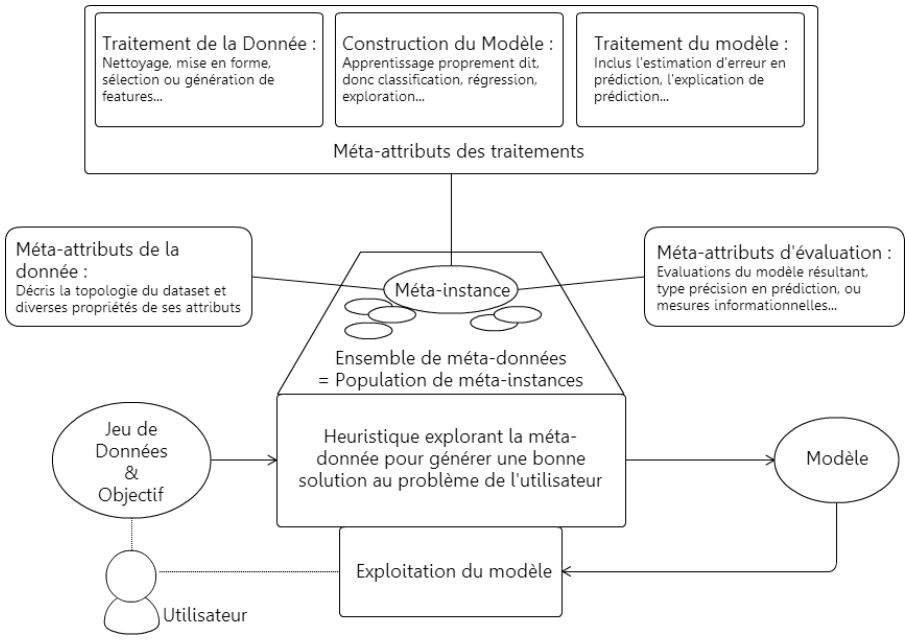


FIG. 1 – *Assistant de modélisation.*

Un élément crucial pour la réalisation de cette approche, ainsi que le sujet de ce papier, sera la caractérisation de ces méta-instances. Ce problème peut être considéré comme une version étendue du problème de caractérisation des jeux de données, étudié par la plupart des approches de méta-apprentissage. Ce dernier consiste en la définition d'un ensemble de propriétés (alors appelées méta-attributs) caractérisant finement des jeux de données, mais néanmoins soumis aux pré-requis imposés par les algorithmes d'apprentissage qui leur seront appliqués. Ces der-

niers imposent notamment l'utilisation de vecteurs de méta-attributs de taille fixe, ce qui implique des agrégations (par exemple la variance individuelle des différents attributs d'un jeu de données devient la variance moyenne sur ce jeu de données...) et donc une importante perte d'information (Kalousis et Hilario, 2001b). Ce problème pourrait être résolu par l'emploi au méta-niveau d'une heuristique évolutionnaire dont la fonction de "*fitness*" reposerait sur une *dissimilarité* entre méta-instances. Cette dissimilarité pouvant être construite sur l'ensemble des informations disponibles, on s'affranchit des restrictions de représentation, se rapprochant en cela des approches "anti-essentialistes" telles que décrites par Duin (2015).

2 Caractérisation des instances d'apprentissage

Cette section concerne les attributs qui décriront nos méta-instances, ainsi appelés méta-attributs. L'ensemble de ces méta-attributs devrait être suffisamment important pour permettre une caractérisation fine de toute expérience de data-mining, mais un équilibre devra être trouvé pour éviter l'abondance de méta-attributs trop dépendants et limiter la complexité computationnelle. De plus, afin de pouvoir discriminer des méta-attributs ou méta-instances, la comparaison de méta-attributs au sein d'une méta-instance, tout comme la comparaison d'un même méta-attribut entre plusieurs méta-instances, devra être possible et sensée.

On peut intuitivement diviser les méta-attributs selon trois dimensions, comme illustré en figure 1. Pour des raisons de volume, on ne s'intéressera ici qu'aux méta-attributs décrivant les jeux de données, ceux décrivant les traitements employés et l'évaluation des modèles résultants seront privilégiés dans de futurs travaux.

Le problème de caractérisation d'un jeu de données a été étudié selon deux axes :

- Le premier consiste en l'emploi de mesures statistiques et information-théorétiques pour décrire le jeu de données. Cette approche, notamment mise en avant par le projet STATLOG (Michie et al., 1994), et employée dans une majorité d'études postérieures (Kalousis, 2002; Vilalta et Drissi, 2002; Leyva et al., 2015), présente nombre de mesures très expressives, mais sa performance repose intégralement sur l'adéquation entre le biais de l'apprentissage effectué au méta-niveau et l'ensemble de mesures choisies. On note parfois l'emploi de techniques de sélection d'attributs à ce méta-niveau (Kalousis et Hilario, 2001a), mais les résultats expérimentaux ne permettent pas de conclure à la supériorité de quelconque mesure indépendamment du méta-apprentissage employé (Todorovski et al., 2000).
- Le second axe d'approche considère quant à lui non pas des propriétés intrinsèques du jeu de données étudié, mais plutôt la performance d'algorithmes d'apprentissage simples exécutés dessus. Introduit comme "*landmarking*" par Pfahringer et al. (2000), cette approche emploie initialement le taux d'erreur d'un ensemble d'algorithmes basiques comme méta-donnée. Comme précédemment, les résultats suggèrent une forte dépendance de l'efficacité de cette approche avec le choix des algorithmes de base et du méta-niveau, ne révélant aucune combinaison uniformément supérieure. Des développements postérieurs ont introduit des mesures plus complexes, tel Peng et al. (2002) proposant comme méta-attributs des propriétés structurelles d'un arbre de décision construit sur la donnée.

Les expériences conduites par Fürnkranz et Petrak (2002) sur ces différentes approches tendent à conclure que toutes peuvent réaliser de bonnes performances dans diverses parties de l'ensemble des jeux de données, sans qu'aucune ne domine globalement.

La transition de paradigme entre notre approche et le méta-apprentissage traditionnel nous affranchit des pertes d'information causées par les agrégations évoquées plus tôt, mais nous place face à un nouveau défi : il est possible de caractériser librement les méta-instances, mais il faut pouvoir les *comparer de manière sensée*. Le coeur du problème réside en la comparaison de méta-attributs décrivant des attributs particuliers de la donnée, comme illustré dans l'exemple suivant.

Exemple *Considérons deux jeux de données, **A** et **B** illustrés en figure 2. **A** décrit 12 attributs de 100 individus, et **B** 10 attributs de 200 individus. On souhaite comparer les résultats de 5 mesures statistiques et informationnelles relevées sur les attributs individuels de ces jeux de données (comme illustré sur le second attribut de **A**). L'information complète que l'on souhaite comparer est donc un vecteur de 60 valeurs pour **A** et de 50 pour **B**.*

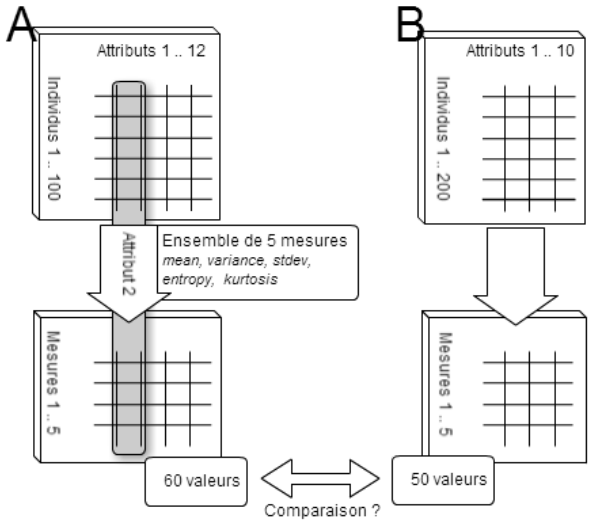


FIG. 2 – *Comparaison de mesures relevées sur des attributs individuels.*

*Notre approche est de comparer les attributs de **A** et **B** par paires les plus similaires, comparant les attributs en surnombre à d'hypothétiques attributs vides. L'hypothèse émise ici est qu'un attribut absent équivaut à un attribut dont aucune valeur n'est connue. Pour en revenir à l'exemple, la comparaison des 5 mesures s'effectuera donc entre l'attribut de **A** et l'attribut de **B** les plus similaires* selon ces mêmes mesures*, puis sur les second plus similaires et ainsi de suite, pour finir par comparer les mesures relevées sur les deux attributs surnuméraires de **A** avec leur valeur sur un hypothétique attribut vide. Cette comparaison par paire permet de s'affranchir de l'ordre de présentation des attributs, qui ne recèle aucune information, se concentrant sur la topologie réelle du jeu de données.*

Supposant qu'une comparaison très expressive résultera en un "*fitness*" plus juste, nous envisageons d'utiliser l'ensemble des mesures, statistiques comme issues de la théorie de l'information, présentées dans la littérature citée précédemment, ainsi que les différents méta-attributs construits sur les approches de "landmarking". Des recherches plus poussées seront cependant nécessaires pour limiter la duplication d'information entre les deux approches.

3 Conclusion et perspectives

Les méta-attributs présentés ici donnent un aperçu de la structure des méta-instances qui seront manipulées par l'heuristique évolutionnaire. En d'autres termes, ces méta-attributs constitueront le génome des méta-instances, dont l'évolution contrainte par divers mécanismes heuristiques devra permettre la découverte des traitements apportant une réponse satisfaisante au besoin de l'utilisateur.

Cependant, afin de compléter et d'évaluer notre approche, plusieurs tâches importantes restent à compléter. Tout d'abord, une représentation du besoin de l'utilisateur permettant son élicitation automatique ou semi-automatique sera nécessaire. En effet, l'assistant s'adressant à des non-experts, il devra aider à la définition formelle du besoin qui indiquera la cible de l'évolution. De plus, la sémantique d'une telle représentation pourra permettre de réduire la complexité computationnelle de l'heuristique en éliminant les méta-instance ne pouvant pas répondre au besoin de l'utilisateur.

La dissimilarité entre méta-instances devra ensuite être formalisée afin de construire la fonction fitness de l'heuristique, qui devra à son tour être définie et calibrée. La création de mécanismes "prédateurs" limitant la prolifération de méta-instances "inutiles" devra également être considérée. On s'intéressera en particulier aux approches génétiques (Bacardit et Llorà, 2013) et mémétiques (Smith, 2007) présentant des propriétés désirables, telles l'absence de contrainte quant à la structure du génome, et un parallélisme natif très souhaitable pour une approche computationnellement très complexe.

Références

Aha, D. W. (1992). Generalizing from case studies : A case study. In *Proc. of the 9th International Conference on Machine Learning*, pp. 1–10.

Bacardit, J. et X. Llorà (2013). Large-scale data mining using genetics-based machine learning. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery 3*(1), 37–61.

Duin, R. P. (2015). The dissimilarity representation for finding universals from particulars by an anti-essentialist approach. *Pattern Recognition Letters 64*, 37 – 43.

Fürnkranz, J. et J. Petrak (2002). Extended data characteristics. Technical report, METAL consortium. Accessed 12/11/15 at `citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.302`.

Gama, J. et P. Brazdil (1995). Characterization of classification algorithms. In *Progress in Artificial Intelligence*, pp. 189–200. Springer.

Kalousis, A. (2002). *Algorithm selection via meta-learning*. Ph. D. thesis, Universite de Geneve.

Kalousis, A. et M. Hilario (2001a). Feature selection for meta-learning. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD '01, London, UK, UK, pp. 222–233. Springer-Verlag.

Kalousis, A. et M. Hilario (2001b). Model selection via meta-learning : a comparative study. *International Journal on Artificial Intelligence Tools 10*(04), 525–554.

Leyva, E., A. Gonzalez, et R. Perez (2015). A set of complexity measures designed for applying meta-learning to instance selection. *Knowledge and Data Engineering, IEEE Transactions on 27*(2), 354–367.

Michie, D., D. J. Spiegelhalter, et C. C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Upper Saddle River, NJ, USA : Ellis Horwood.

Mitchell, T. M. (1997). Machine learning. *Burr Ridge, IL : McGraw Hill 45*.

Peng, Y., P. A. Flach, P. Brazdil, et C. Soares (2002). Decision tree-based data characterization for meta-learning. Accessed 12/11/2015 at `citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.4890`.

Pfahringer, B., H. Bensusan, et C. Giraud-Carrier (2000). Tell me who can learn you and i can tell you who you are : Landmarking various learning algorithms. In *Proceedings of the 17th international conference on machine learning*, pp. 743–750.

Raynaut, W., C. Soule-Dupuy, et N. Valles-Parlangeau (2015). Addressing the meta-learning problem with metaheuristics (extended abstract). In *Metaheuristics International Conference*.

Smith, J. E. (2007). Coevolving memetic algorithms : a review and progress report. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on 37*(1), 6–17.

Todorovski, L., P. Brazdil, et C. Soares (2000). Report on the experiments with feature selection in meta-level learning. In *Proceedings of the PKDD-00 workshop on data mining, decision support, meta-learning and ILP : forum for practical problem presentation and prospective solutions*, pp. 27–39. Citeseer.

Vilalta, R. et Y. Drissi (2002). A perspective view and survey of meta-learning. *Artif. Intell. Rev. 18*(2), 77–95.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation 8*(7), 1341–1390.

Xu, L., F. Hutter, J. Shen, H. H. Hoos, et K. Leyton-Brown (2012). Satzilla2012 : improved algorithm selection based on cost-sensitive classification models. *Balint et al.*, 57–58.

Summary

Machine learning has proven to be a powerful tool in diverse fields, and is getting more and more widely used by non-experts. One of the foremost difficulties they encounter lies in the choice and calibration of the machine learning algorithm to use. Our objective is thus to provide assistance in the matter, using a meta-learning approach based on an evolutionary heuristic. We expand here previous work presenting the intended workflow of a modeling assistant by describing and discussing the characterization of learning instances we intend to use.

Apprentissage du signal prix de l'électricité : prédictions à long terme

Louis-Victor PASQUIER* Antoine CORNUEJOLS* Suzanne PINSON**

*UMR MIA-Paris, Agroparistech, INRA, Université Paris-Saclay 75005, Paris, France
**Université Paris-Dauphine, LAMSADE, CNRS-UMR 7243, PSL 75775 Paris Cedex 16

Résumé. Aujourd'hui, l'électricité constitue une marchandise comme une autre et le prix de l'électricité est sujet à de fortes variations pouvant doubler voire tripler d'une heure à l'autre. De plus, le caractère non stockable de l'électricité rend le comportement de son prix plus volatile encore. Le fait que la demande en électricité dépende de surcroît de la météo (température, vitesse du vent, préecipitations) ainsi que des activités humaines (week-end, congés ...) ajoute encore à la difficulté pour appréhender les tarifs. Contrairement aux modèles existants peu performants sur les prévisions à long terme, nous proposons un nouveau modèle fondé sur l'apprentissage d'arbres de régression qui permet une interprétation des résultats sur le long terme à un faible coût calculatoire. Nous validons notre approche sur les séries temporelles échantillonnées recueillies auprès d'un fournisseur d'électricité.

1 Introduction

L'électricité est devenue une matière première comme une autre, et la prédiction de son prix à court, moyen et long terme devient un enjeu très important pour des pans entiers de l'économie. Cependant, la prédiction de ce prix s'apparente à un problème de régression complexe dans lequel interviennent de nombreux facteurs et des cycles multiples. Comment dès lors faire des prédictions fiables, particulièrement à long terme ?

Plus spécifiquement, le problème étudié est celui de la prédiction d'un prix (un réel) sur une durée plus ou moins longue et à terme plus ou moins rapproché, à partir d'une série temporelle multi-variée comprenant des informations sur l'évolution des prix, mais aussi potentiellement sur des données extrinsèques telles que la météo.

Dans cet article, après un état de l'art aussi extensif que possible sur la prédiction des tarifs et des consommations électriques (section 2), nous présentons (section 4) une méthode utilisant du boosting sur arbres de régression dédiée à des séries temporelles de longue durée. Testée sur des données réelles portant sur une durée d'un an, cette méthode donne des résultats excellents (section 5). Nous en analysons les raisons et les perspectives d'utilisation en conclusion.

2 État de l'art

De nombreuses approches ont été essayées, basées sur des modélisations multi-agents, ou sur des modélisations stochastiques. Des méthodes inductives de nature statistique ou issues de l'apprentissage automatique ont également été proposées. Nous les décrivons rapidement ici. Globalement, elles n'ont pas apporté de solutions satisfaisantes à tous les desiderata des utilisateurs : fiabilité de prédiction, spécialement à long terme, mise en évidence des variables explicatives importantes et complexité calculatoire maîtrisée.

2.1 Modèles multi-agents

Ces méthodes sont basées sur l'analyse ou la simulation du comportement de différents agents suivant des règles simples. On distingue les modèles de recherche d'un équilibre et les modèles de simulations.

— **Modèles de recherche d'équilibre**. Ces modèles statiques s'appuient sur une définition de l'équilibre traduite par des équations algébriques ou différentielles. Si ces équations possèdent une solution, il est souvent très délicat de parvenir à l'identifier et il faut alors recourir à des heuristiques. De plus, il est difficile de représenter les interactions entres les différents compétiteurs et leur comportement.

— **Utilisation de systèmes multi-agents**. Les modèles de simulation multi-agents permettent une expression plus facile des interactions et peuvent converger vers une solution ou un comportement général relativement rapidement. De ce fait, depuis les années 1995, les systèmes multi-agents ont largement été acceptés dans la résolution de problèmes à la fois théoriques et pratiques du marché des énergies (Guerci et al. (2010)). Ils conviennent pour identifier des comportements généraux qui émergent des diverses interactions entre les agents, mais moins pour prédire précisément des prix.

Mais si les modèles multi-agents sont par construction extrêmement flexibles et évolutifs, et capables de modéliser des réalités particulièrement complexes à travers des populations d'individus simples et autonomes, le degré de liberté dans la construction de tels modèles est si étendu qu'il est parfois difficile d'identifier quels sont les acteurs pertinents qui contribuent véritablement à la détermination du prix de l'électricité. De plus, de nombreuses hypothèses sont souvent posées sans base solide dans ces modèles.

2.2 Modèles fondamentaux

Cette classe de modèles tente de représenter les liens entre la physique et l'économie. Les variables de ces modèles sont typiquement des grandeurs physiques telles que la température, la capacité disponible, la force du vent ... La plupart du temps ces modèles sont intégrés avec d'autres modèles de type prédictif (e.g. réseau de neurone, régression ...). Deux catégories peuvent être distinguées :

1. Les modèles riches en paramètres ou complets
2. Les modèles parcimonieux

1. **Les modèles complets**. Un exemple est donné par l'article de Johnsen (2001) qui a élaboré un modèle tenant compte des précipitations, des chutes de neige, et des températures extérieures pour expliquer le prix de l'électricité. De même, les travaux de

Vahviläinen et Pyykkönen (2005) ont démontré que l'on pouvait retrouver un prix à l'aide de 13 paramètres climatiques, 4 paramètres relatifs à la demande et 10 paramètres relatifs à l'offre, et en reliant ces 27 paramètres par un jeu 29 équations. Leur système semble cohérent jusqu'à un horizon de prédiction d'un mois.

2. **Les modèles parcimonieux**. Coulon et Howison (2009) ont développé un exemple de modèle fondamental pour le prix de l'électricité. Les facteurs sous-jacents du modèle s'appuient sur un processus stochastique (ces facteurs correspondent au prix des combustibles, à la demande en énergie et à la capacité de production électrique). Par ailleurs, une fonction paramétrique de l'offre en énergie associe à un ensemble de facteurs explicatifs un prix pour l'énergie correspondante.

Ces modèles nécessitent que des historiques longs et finement échantillonnés soient disponibles. Ils reposent souvent sur des hypothèses fortes concernant les liens entre les phénomènes physiques et économiques (linéarité, indépendance ...). Les prix résultants de tels modèles peuvent donc être fortement biaisés. Enfin, plus le modèle est complexe et fait intervenir de paramètres, plus il est difficile d'ajuster ces paramètres.

2.3 Modèles de formes réduites

Ces modèles incluent principalement les modèle dits à « saut-diffusion » et les modèles de régime de Markov. Ils reposent sur des équations stochastiques gouvernant les variations de prix en fonction de différents facteurs (modèles saut-diffusion (Geman et Roncoroni (2006)) ou bien sur des modélisations par transitions aléatoires entre régimes différents (modèles markoviens (Karakatsani et Bunn (2010)).

Ces modèles n'ont pas vocation à précisément prédire les prix horaires de l'électricité. Cependant, le comportement global du signal prix est souvent recherché dans ces modèles. Et ils se sont avérés assez performants vis à vis de la prévision des pics de prix (Weron, 2014).

2.4 Modèles statistiques

Les modèles statistiques combinent différentes variables pour rendre compte d'un phénomène. Typiquement, ces variables peuvent être constituées de l'historique passé de la variable à expliquer ou d'autres variables (qualifiées d'exogènes). Ainsi, un historique des prix de l'électricité accompagné d'un historique des disponibilités en carburants fossiles en énergies renouvelables peut permettre d'obtenir une prédiction des prix à venir.

On distingue deux grandes catégories de modèles :

1. les modèles additifs
2. les modèles multiplicatifs

Parmi les nombreuses méthodes développées, on peut noter :

— **Jours similaires et lissage exponentiel.** Cette méthode s'appuie sur la ressemblance entre les jours passés et les jours à prédire. Il est intuitif de considérer que les prix de deux jours proches (d'après des caractéristiques à définir telle que la saison, le mois, le jour ...) ont une grande chance d'être proches ou de présenter un comportement similaire (Shahidehpour et al. (2002); Weron (2006); Conejo et al. (2005)).

La méthode que nous proposons présente de nombreuses similitudes avec cette approche.

— **Modèles de régression**. Dans sa forme classique, la régression relie linéairement les variables entre elles :

$$P_t = BX_t + \epsilon_t = b_1 X_t^{(1)} + ... + b_k X_t^{(k)} + \epsilon_t \quad (1)$$

où B est un vecteur $(1 \times k)$ de coefficients constants, X_t un vecteur $(k \times 1)$ de variables explicatives et P_t est le prix a prédire à l'instant t.
Il existe des régressions temporelles (ou TRV pour *time varying regression*) qui permettent au coefficient associé à chaque variables explicatives de varier de façon continue avec le temps. On a alors la relation (1) modifiée en (2) :

$$P_t = B_t X_t + \epsilon_t = b_{1,t} X_t^{(1)} + ... + b_{k,t} X_t^{(k)} + \epsilon_t \quad (2)$$

Dans le cas où les variables explicatives sont des valeurs passées de la variable à expliquer, on parle de modèles *auto-régressifs* (les modèles dits *AR*). Dans le cas où il s'agit d'autres grandeurs, on parle de modèles *exogènes* (les modèles dits *X*). Il existe naturellement des modèles combinant les deux types de variables (les modèles *ARX*).
On retrouve dans ces familles les modèles ARIMA (Auto-Regressive Integrated Moving Average) et dérivés comme SARIMA (le "S" pour *Seasonal*) qui traitent particulièrement les variables sujettes à une certaine périodicité. C'est notamment le cas du prix de l'électricité. Ces modèles intègrent également une composante MA semblable au lissage exponentiel qui vise en général à accorder plus d'importance à un historique récent.
En s'appuyant sur les fondations de ces modèles principaux, de nombreux chercheurs ont modifié leur structure afin de répondre à des attentes particulières. En règle générale, ces spécifications de modèle vont dans le sens de la complexification du modèle (Cuaresma et al. (2004); Weron et Misiorek (2005)).

Les méthodes statistiques prédisent le prix de l'électricité en s'appuyant sur des combinaisons mathématiques des prix précédents et/ou de facteurs exogènes. Ainsi, la précision du modèle et sa pertinence ne dépendent pas uniquement de la qualité intrinsèque du modèle mais aussi de la capacité à lui injecter des données appropriées et précises en entrée. Il n'est pas aisé d'obtenir des historiques précis à une fine résolution sur un horizon lointain. Autant ces données restent relativement accessibles pour la variable à prédire (le prix de l'électricité), autant, il est beaucoup plus épineux d'obtenir ces valeurs pour les stocks de combustibles, la météo ou encore un historique de l'appel de charge des consommateurs. Les méthodes statistiques se révèlent généralement nettement moins performantes que d'autres méthodes (en particulier stochastiques) en ce qui concerne la détection et la simulation des pics de prix.

2.5 Méthodes d'apprentissage automatique

Afin de s'adapter au caractère hautement non linéaire des phénomènes étudiés, des techniques d'apprentissage automatique ont été utilisées dans de nombreux travaux. Ce sont en particulier les réseaux de neurones (Mandal et al. (2006); Pao (2006); Zhang et al. (2009);

Catalao et al. (2009)), et parmi ceux-ci, particulièrement les réseaux ne neurones récurrents, et les Séparateurs à Vastes Marges (SVM) dotés de fonctions noyaux appropriées.

L'usage de ces techniques demande cependant une bonne familiarité et peuvent être difficiles à paramétrer correctement. Globalement, les meilleures performances sont observées pour des prédictions à court terme.

Ce papier propose une méthode pour apprendre le modèle d'une année type complète. L'approche utilisée se rapproche par certains points de la méthode d'identification des jours similaires. Elle s'appuie sur un apprentissage par arbres de régression en raison de leur versatilité et de leur capacité expressive. Nous avons cherché également jusqu'à quel point il est possible de limiter les données à fournir en entrée. Nous montrons qu'avec la donnée de la date seulement il est possible d'expliquer une part significative de la variance du prix de l'électricité. La méthode présentée est ainsi à la fois puissante et aisée à mettre en œuvre.

Après avoir présenté le jeu de données, nous montrons comment les arbres de régression peuvent aider à mieux comprendre les phénomènes qui régissent les variations du prix de l'électricité.

3 Le type de données

Le signal que nous chercherons à analyser est le prix spot de l'électricité (c'est à dire payé dans un marché au comptant). Ce dernier s'exprime en *€/MWh*.

Les données se présentent sous la forme d'une série temporelle échantillonnée au pas de temps horaire. L'historique des prix s'étend sur une année complète, soit un total d'environ 100 000 points relevés.

Afin de normaliser les données, le prix d'une heure de consommation est exprimé par rapport au prix moyen de l'année étudiée.

$$y_{date_k,annee_j} = \frac{Prix_{date_k,annee_j}}{\frac{1}{n_j}\sum_i Prix_{date_i,annee_j}} \tag{3}$$

avec :

- y la variable (prix) à prédire
- $date_j$ l'année considérée
- $date_k, anne_j$ la date (jour et heure) et année du point relevé
- n_j le nombre de points relevés au cours de l'année j considérée

Transformation de la date en attributs Les formats de date du type AAAA-MM-JJ hh :mm sont décrits gâce aux attributs suivants permettant une représentation numérique :

- No du jour de l'année (dans [1 :366])
- Mois dans l'année (dans [1 :12])
- Position dans la semaine (dans [1 :7])
- Heure de la journée (dans [0 :23])
- Booléen de week-end (dans 0,1)
- Booléen de jour férié (dans 0,1)

La figure 1 montre un échantillon des donnés utilisées.

	date	dayofyear	month	dayofweek	hour	we	hol	target
1	2005-01-01 00:00:00	1	1	5	0	1	1	0.3181746
2	2005-01-01 01:00:00	1	1	5	1	1	1	0.3813810
3	2005-01-01 02:00:00	1	1	5	2	1	1	0.3218170
4	2005-01-01 03:00:00	1	1	5	3	1	1	0.2339708
5	2005-01-01 04:00:00	1	1	5	4	1	1	0.1759066
6	2005-01-01 05:00:00	1	1	5	5	1	1	0.1767637
7	2005-01-01 06:00:00	1	1	5	6	1	1	0.1765494
8	2005-01-01 07:00:00	1	1	5	7	1	1	0.1754781
9	2005-01-01 08:00:00	1	1	5	8	1	1	0.1748353
10	2005-01-01 09:00:00	1	1	5	9	1	1	0.1763351
11	2005-01-01 10:00:00	1	1	5	10	1	1	0.2571108

FIG. 1 – *Échantillon de l'ensemble d'apprentissage (les températures n'apparaissent pas puisqu'elles ne sont pas exploitées. La colonne target est explicité dans l'équation 3).*

4 La méthode d'apprentissage

La méthode d'apprentissage utilisée est celle des *Random Forests* (Breiman, 2001) utilisant des arbres de régression comme régresseurs faibles. Chacun des régresseurs apprend des régularités simples sur le signal temporel. L'objectif est de considérer un grand nombre de régularités à travers différents arbres et d'effectuer une moyenne pondérée des prédictions de chacun des arbres pour déterminer la variable d'intérêt y.

Le principe des random forests est de forcer la diversité des régularitées apprises (arbres), en tirant aléatoirement à chaque itération le sous-échantillon d'apprentissage utilisé, ici d'une taille de 63% de l'ensemble d'apprentissage par tirage avec remise, et en contraignant l'apprentissage de chaque arbre à n'utiliser qu'un sous-ensenble tiré aléatoirement d'attributs.

Les paramètres de la méthode sont :

— le nombre d'attributs tirés aléatoirement pour constituer le sous-espace de redescritpion
— le nombre d'arbres à retenir
— la profondeur limite des arbres
— le nombre minimal d'exemples par feuille

5 Résultats expérimentaux

Pour évaluer la performance des modèles, nous nous sommes appuyé sur la métrique de la variance expliquée (part de variance réduite par le modèle mis en place) qui est une forme de l'erreur quadratique moyenne normalisée.

$$Var.exp. = 1 - \frac{Variance(y_{observée} - y_{modèle})}{Variance(y_{observée})} \tag{4}$$

Le tableau 1 récapitule les résultats de différents paramétrages pour l'apprentissage.

n	Ens. app	Ens. test	Nb arbres	Nb attr.	Min. ex./feuil.	Var. exp. app.	Var. exp. test
1	2005-2012	2013-2014	200	5	50	82%	58%
2	2005-2012	2014	200	5	50	82%	62%
3	2010-2012	2014	200	5	50	83%	61%
4	2010-2012	2014	500	5	50	83%	61%
5	2010-2012	2014	200	1	50	60%	45%
6	2010-2012	2014	200	3	50	82%	60%
7	2010-2012	2014	200	5	1	94%	55%
8	2010-2012	2014	200	5	100	78%	60%
9	2005-2012	2014	1	6	1	90%	6%
10	2005-2012	2014	200	5	1	76%	62%

TAB. 1 – *Tableau récapitulatif des paramétrages.*

Une variance expliquée de $\approx 0,6$ correspond à un bon degré de prédictivité des variables d'entrée pour prédire la variable de sortie y. C'est à peu près le degré obtenu avec les modèles appris par la méthode de random forest sur les variables explicatives choisies.

L'écart de performance entre les lignes 1 & 2 s'explique par la rudesse de l'hiver 2013, particulièrement long avec des températures très basses jusqu'en avril. C'est pourquoi l'année 2013, atypique, est exclue de l'ensemble d'apprentissagepour les lignes suivantes. Le tableau montre également que l'augmentation du nombre de régresseurs n'améliore pas nécessairement la qualité du modèle (ligne 4) : il n'y a pas de sur-apprentissage, mais pas non plus d'amélioration de la performance en test. On notera que l'apprentissage par régresseurs à 1 nœud (ligne 5) conduit à de mauvaises performances, mais que l'utilisation de 3 attributs par régresseur (ligne 6) est à peine moins bonne qu'avec plus d'attributs, ce qui est important si l'on tente d'interpréter le modèle appris. En revanche, quand tous les attributs sont permis pour chaque régresseur, un phénomène de sur-apprentissage extrême survient (ligne 9).

La variance expliquée est un critère objectif important pour mesurer le degré prédictif des variables utilisées, mais cela ne suffit pas à caractériser la valeur d'un modèle de prévision temporelle. Nous étudions dans la suite les caractéristiques du signal prix retranscrites par le modèle sur différentes échelles de temps (journalière, hebdomadaire, annuelle)

5.1 Comportement sur une journée

Le modèle appris pour une journée (figure 2) prédit bien les heures les moins chères (5h la nuit et 15h en journée) et les heures les plus chères (8h et 18h). La figure 2 reflète aussi les limites du modèle. En effet, les amplitudes absolues des pics et des creux du modèle sont généralement moins prononcés par rapport à la réalité. Cependant, pour les décideurs intéressés par les dates (ici en heures et minutes) auxquelles surviennent les variations importantes, et non pas nécessairement par les amplitudes, il est remarquable que le modèle appris soit parfaitement synchrone avec la réalité.

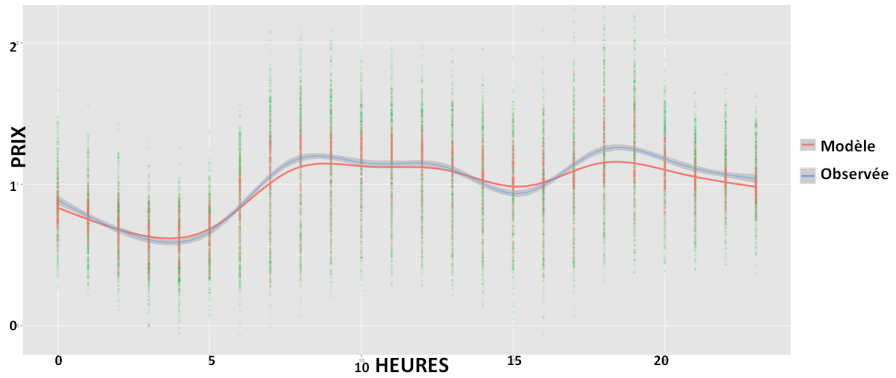


FIG. 2 – *Comportement du modèle (courbe en rouge) sur un jour moyen de l'année. L'ensemble de test est représenté par la courbe en bleu (année 2014).*

5.2 Comportement sur une semaine

Le modèle appris pour une semaine type retrouve bien l'alternance des jours (forte demande) et des nuits (faible demande) sur la figure 3. Par ailleurs, il prédit bien une amplitude de variation élevée en semaine et moindre en week-end, tandis qu'est mise en évidence une différence significative entre les samedis (prix intermédiaire) et les dimanches (prix les plus faible).

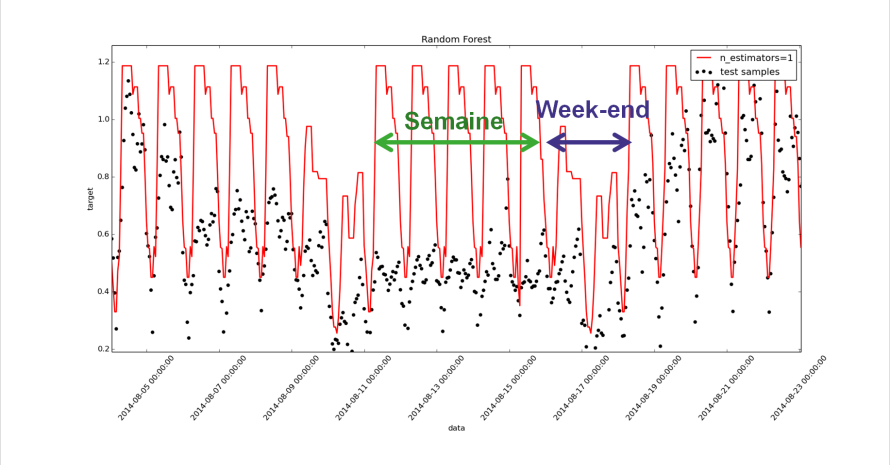


FIG. 3 – *Prédiction du modèle (courbe en rouge) contre les points de l'ensemble de test (points en noir) pour une semaine type.*

La semaine sur laquelle est centré le graphe 3 n'est pas correctement retranscrite en ampli-

tude de variation par la modèle. Il s'agit en effet d'une semaine exceptionnelle de l'année 2014 qui est difficile à apprendre sans tomber par ailleurs dans une situation de sur-apprentissage.

Cependant, des comportements particuliers, tels que les jours fériés, font partie des régularités qui intéressent les décideurs qui veulent tirer parti des baisses de tarif de l'électricité. La figure 4 témoigne que le modèle est capable de rendre compte des jours fériés. On retrouve en effet un baisse significative du prix lors du 1er et du 8 mai, corroborée par les observations réelles. Les jours fériés se comportent en général comme des week-end (fermeture des commerces et du tertiaire ...)

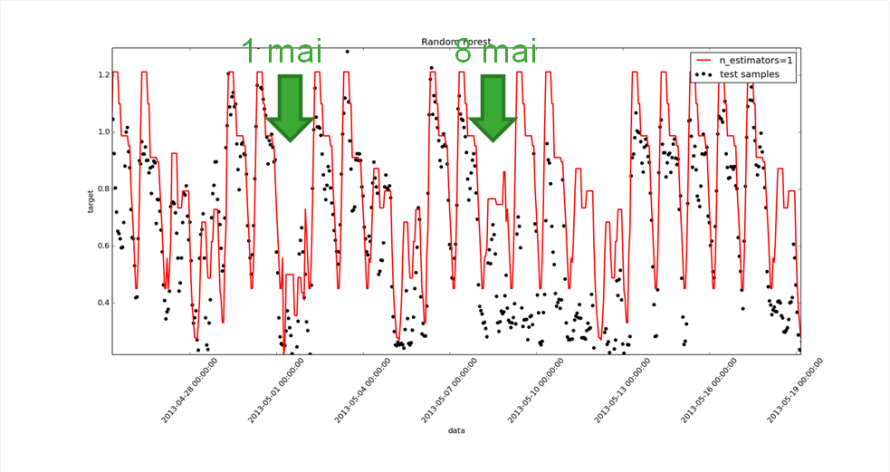


FIG. 4 – *Comportement du modèle (courbe en rouge) sur la semaine du 1er mai et du 8 mai. L'ensemble de test est représenté par les points en noir.*

Les figures 5 et 6 rassemblent les comportements moyens prédits par le modèle appris des prix de l'électricité suivant les jours de semaine, respectivement en hiver et en été. Ici aussi le modèle est conforme aux connaissances des experts qui s'attendent à ce que :

— les pointes (notamment celle de 18h) soient de plus forte amplitude de variation en hiver qu'en été. (Environ 1 fois le $Prix_{moyen}$ en été contre 1.8 fois en hiver).
— les pointes d'été soient plus étalées dans le temps que les pointes d'hiver.

5.3 Comportement sur une année

Les résultats publiés dans la littérature sur le sujet montrent que la grande majorité des méthodes d'apprentissage ont du mal à faire des prédictions correctes à l'échelle de l'année pour la prédiction du tarif de l'électricité. La figure 7 concernant une année complète (2014) montre l'accord des prédictions du modèle avec la réalité dans les grandes variations annuelles. On remarque ainsi, par exemple, que la forte chute du prix de l'électricité fin décembre est bien prédite, qui correspond aux départs massifs en congé et à la libération de l'essentiel des bâtiments de bureaux à cette période.

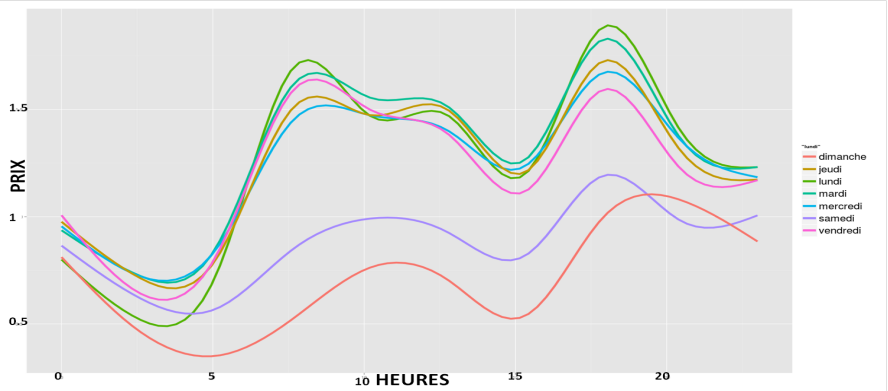


FIG. 5 – *Comportement moyen du modèle selon les différents jours d'une semaine d'hiver.*

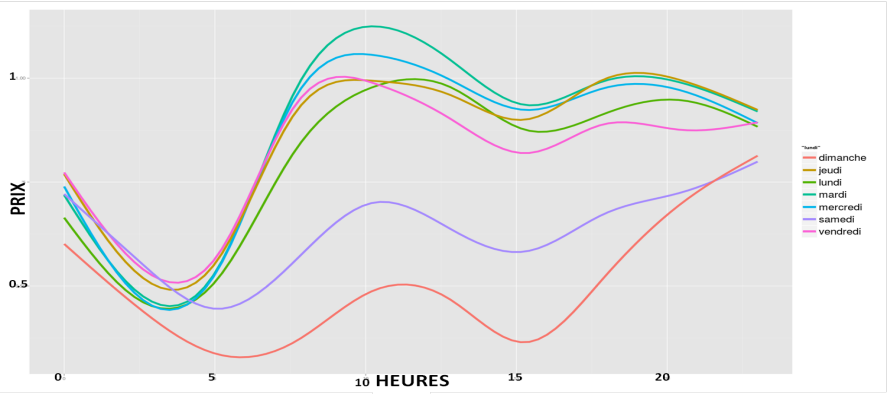


FIG. 6 – *Comportement moyen du modèle selon les différents jours d'une semaine d'été.*

6 Conclusion

La prédiction sur les séries temporelles est un problème qui prend une importance grandissante en raison des bases de données nouvellement disponibles et du champ de plus en plus étendu des applications possibles. Pour certaines d'entre elles, comme la prédiction du tarif de l'électricité, il est essentiel de pouvoir faire des prédictions à court, moyen et long terme, alors même que les facteurs à prendre en jeu sont variés, ici la saison, la météo, les habitudes sociales, etc. De plus, il est souvent important de pouvoir mesurer l'importance de ces variables pour la prédiction. Il faut donc à la fois des méthodes de prédiction performantes à divers horizon temporels, et qui produisent des modèles qui soient interprétables.

Dans le domaine de la prédiction des tarifs électriques, les méthodes les plus répandues, à base de modèles auto-régressifs, ne permettent que des prédictions à court terme. Et les autres

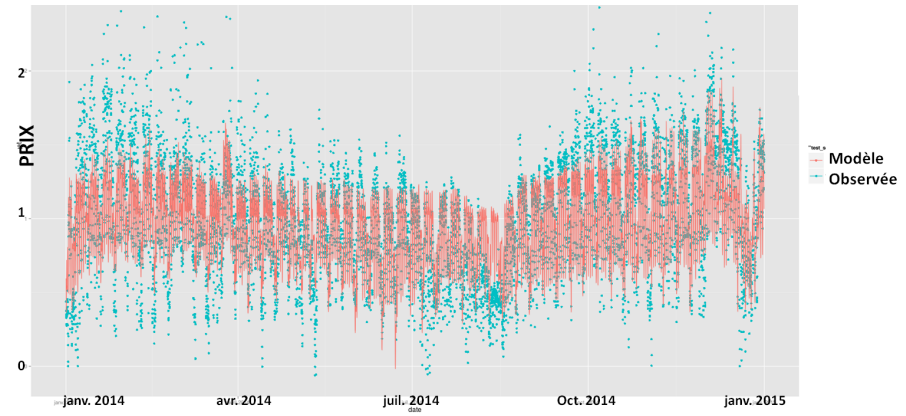


FIG. 7 – *Modèle annuel (en rouge) contre l'ensemble de test (2014 en bleu).*

méthodes, multi-agents, à base de modèles physiques, ou par technique de réseaux de neurones ou de SVM, souvent également peu performantes en prédiction à long terme, sont de plus très peu interprétables.

Afin de satisfaire à la fois les demandes en termes de prédiction et d'interprétabilité du modèle appris, nous avons présenté une technique qui utilise la méthode des random forests utilisant des arbres de régression comme méthodes de base. En s'appuyant sur les descripteurs de date seulement, le modèle appris retranscrit la multi-saisonnalité ainsi que des événements ponctuels tels que les jours fériés ou encore (fig. 7) le départ en congé massif de fin décembre caractérisés par une forte baisse du prix de l'électricité.

La méthode présentée s'applique directement à toutes les grandeurs variant pseudo-périodiquement au cours du temps en fonction de nombreuses variables à valeurs numériques ou traduites en valeurs numériques comme nous l'avons fait pour la date. De plus, la méthode des random forests est peu coûteuse à mettre en œuvre en terme de calculs.

Références

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Catalao, J., S. Mariano, V. Mendes, et L. Ferreira (2009). Short-term electricity prices forecasting in a competitive market : a neural network approach. *Electric Power Syst. Res.*, 1297–1304.

Conejo, A., J. Conferas, R. Espinola, et M. Plazas (2005). Forecasting electricity prices for a day-ahead pool-based electric energy market. *International journal of forecasting*, 435–462.

Coulon, M. et S. Howison (2009). Stochastic behavior of the electricity bid stack : from fundamental drivers to power prices. *Journal of Energy Markets 2*(1), 29–69.

Cuaresma, J., J. Hlouskova, S. Kossmeier, et M. Obersteiner (2004). Forecasting electricity spot prices using linear univariate time-series models. *Applied Energy*, 87–106.

Geman, H. et A. Roncoroni (2006). Understanding the fine structure of electricity prices. *Journal of Business*, 1225–1261.

Guerci, E., M. Rastegar, et S. Cincotti (2010). Agent-based modeling and simulation of competitive wholesale electricity market clearing price prediction. *Handbook of power system II - Energy systems*, 241–286.

Johnsen, T. (2001). Demand, generation and price in the norwegian market of electric power. *Energy Economics*, 227–251.

Karakatsani, N. et D. Bunn (2010). Fundamental and behavioural drivers of electricity price volatility. *Studies in Nonlinear Dynamics and Econometrics*.

Mandal, P., T. Senjyu, et T. Funabashi (2006). Neural networks approach to forecast several hour ahead electricity prices and loads in deregulated market. *Energy Conversion and Management*, 2128–2142.

Pao, H. (2006). A neural network approach to m-daily-ahead electricity price prediction. *Lecture Notes in Computer Science*, 1284–1289.

Shahidehpour, M., H. Yamin, et Z. Li (2002). Market operations in electric power systems : forecasting, scheduling, and risk management. *Wiley*.

Vahviläinen, I. et T. Pyykkönen (2005). Stochastic factor model for electricity spot price - the case of nordic market. *Energy Economics*, 351–367.

Weron, R. (2006). Modeling and forecasting electricity loads and prices : a statistical approach. *Chichester : Wiley*.

Weron, R. (2014). Electricity price forecasting : A review of the state-of-the-art with a look into the future. *International Journal of Forecasting 30*(4), 1030–1081.

Weron, R. et A. Misiorek (2005). Forecasting spot electricity prices with time series models. *IEEE Conference Proceedings EEM05*, 133–141.

Zhang, G., B. Patuwo, et M. Hu (2009). Electricity price forecasting in deregulated markets : a review and evaluation. *International journal of Electrical Power and Energy Systems*, 13–22.

Summary

Nowadays, electricity is treated and traded as an ordinary good. The price of electricity is subject to large variations with a price that can double or even triple within an hour. Moreover, electricity is an energy that cannot be stored. Consequently, its price is even more volatile, and a smart and careful management is needed to balance supply and demand in order to avoid critical situation as black-out. The fact that demand also depends on weather (temperature, wind speed, rainfall ..) as well as human behavior (weekend, holidays ...) makes the understanding of electricity price even more complex. In order to overcome these issues, lots of statistical technics have been suggested. However, most of them are suitable only for short-term forecasting when a long-term vision is often desirable. In this paper, we introduce a new regression tree based model that allows an easy interpretation of results on a long-term period. This model has the extra benefit to require few computational resources. Our approach is validated by temporal series collected from an electricity provider.

Représentation de données temporelles par un modèle à échelle logarithmique

Brieuc Conan-Guez*, Alain Gély*, Lydia Boudjeloud-Assala*, Alexandre Blansché*,
Dario Compagno**, Arnaud Mercier***

*LITA, université de Lorraine,
{brieuc.conan-guez, alain.gely, lydia.boudjeloud-assala, alexandre.blansche}@univ-lorraine.fr
http://www.lita.univ-lorraine.fr/
**CIM, Université Sorbonne Nouvelle - Paris 3
***CARISM, université Panthéon-Assas - Paris 2

Résumé. Dans ce travail, nous nous intéressons aux tweets associés à un événement source (publication d'articles en ligne), et nous représentons la distribution temporelle des tweets par un modèle à résolution variable : haute résolution au début de la série temporelle et plus grossière à la fin. À cet effet, nous proposons d'utiliser un modèle à noyaux basé sur une transformation logarithmique des données.

1 Introduction

Dans ce travail, nous nous intéressons à des familles de séries temporelles particulières, initiées par la survenue d'un événement « source ». Il ne s'agit donc pas d'étudier des séries temporelles sur une fenêtre arbitraire (un mois, un an, etc.) mais d'observer la survenue d'un ensemble d'événements en réponse à l'événement initial. Ce travail a d'abord été motivé par l'étude de la dissémination d'information sur les réseaux sociaux. Plus particulièrement, on s'intéressera à la vie d'un article de presse sur Twitter. L'événement source, la publication d'un article en ligne, étant suivie par une série temporelle d'autres événements : les tweets mentionnant cet article de presse. Cependant, ce travail est transposable à de nombreux cas. On pourra par exemple penser à la dynamique des échanges sur une liste de diffusion (l'événement initial étant la diffusion d'un message polémique, les événements suivants les réponses et commentaires sur ce message) ou encore aux occurrences d'interventions d'équipes de secours (police, samu, etc.) après une crise. (Chae et al., 2014) ont d'ailleurs étudié les occurrences de tweets en réponse à une catastrophe naturelle.

Tous ces exemples ont pour point commun d'avoir un événement initial (l'événement source) qui va engendrer une succession d'autres événements (événements réponses). Dans tous les cas, on peut penser que l'observation des événements réponses proches de la source (de l'ordre de la minute) n'aura pas le même poids dans l'analyse que celle des événements lointains (de l'ordre du mois).

Afin de chercher à dégager les grandes familles de comportements de partage d'articles, nous nous proposons d'utiliser un algorithme de classification non supervisée. Pour ce faire,

il faut donc trouver une représentation des données qui capture l'information finement lorsque l'on est proche de l'événement initial, tout en synthétisant l'information pour les survenues plus tardives, et cela de façon à avoir une description simple et compacte (vecteur de description) de la série temporelle.

Nous proposons une transformation des données de type logarithmique de façon à avoir un niveau de détail d'autant plus élevé que les événements sont proches de leurs sources, tout en permettant de garder trace des événements réponses plus lointains. La transformation logarithmique des données est une approche classique, qui est notamment évoquée dans (Silverman, 1986), quant au modèle à noyaux que nous nous proposons d'utiliser pour modéliser la distribution temporelle des événements réponses, ses propriétés théoriques sont étudiées en détail dans (Charpentier et Flachaire, 2014).

2 Représentation des données

Dans ce travail, nous nous intéressons à un corpus d'articles publiés entre avril et septembre 2014 (soit sur une durée de 5 mois) dans trois media différents : le Monde, le Figaro et Libération. À partir d'avril 2014 et ceci pendant 6 mois (soit jusqu'à un mois après la publication du dernier article), l'ensemble des tweets se rapportant à ces articles a été collecté. Cette durée de collecte assure que pour chaque article, nous disposons d'au moins un mois de données (pour les articles parus fin septembre), et d'au plus 6 mois (pour les articles parus au début de la collecte). Comme nous ne souhaitons analyser que les articles ayant eu un fort impact sur Twitter, les articles pour lesquels le nombre de tweets est inférieur à 100 ont été écartés de l'étude. Le corpus compte donc finalement 6 177 articles pour un total de 822 775 tweets.

Nous disposons de la date et de l'heure de la publication de chaque article ainsi que de celles de chaque tweet. Afin d'étudier les usages de publication de tweets une fois l'article paru, nous nous intéresserons dans la suite de l'étude non pas à la date des tweets, mais à l'intervalle de temps entre la date de publication des articles et la date de publication des tweets associés. On pose T^j la variable aléatoire modélisant la distribution temporelle des tweets se rapportant à l'article j. Chaque article est donc représenté par un vecteur de k_j réalisations $(t_1^j, \ldots, t_i^j, \ldots, t_{k_j}^j)$ de T^j, où k_j est le nombre de tweets se rapportant à l'article j. La valeur de k_j est dans la pratique comprise entre 100 et 200.

Comme indiqué dans l'introduction, nous souhaitons obtenir une représentation de chaque article qui réponde aux deux exigences suivantes :

— la représentation doit être compacte, i.e. chaque article ne doit être décrit que par un petit nombre de valeurs (de l'ordre d'une vingtaine). Ce qui facilitera premièrement les interprétations, et autorisera l'utilisation de méthodes dont le coût est très dépendant de la dimension des données (calcul de la déformation temporelle dynamique (DTW) entre deux séries par exemple) ;

— l'autre exigence à laquelle nous devons répondre nous est imposée par la nature même du problème : nous souhaitons garder un haut niveau de détail lors des premiers instants suivant la parution de l'article (résolution d'analyse de l'ordre de la minute), alors qu'une description relativement grossière est souhaitée dés lors que nous nous intéressons à des instants suffisamment éloignés de la date de parution de l'article (résolution d'analyse de l'ordre du mois pour la fin des séries temporelles).

Afin de répondre à ces deux exigences, nous nous proposons de représenter la distribution temporelle des tweets de chaque article par un modèle à noyaux à échelle variable : le modèle autorisera une représentation fine (resp. grossière) du début (resp. fin) de la série temporelle.

Le modèle que nous utilisons dans cette étude a été proposé dans (Charpentier et Flachaire, 2014) pour la modélisation de distributions à queue non exponentiellement bornée (*heavy-tailed distribution*). Les auteurs citent comme exemple de telles distributions la distribution des revenus des ménages au Royaume-Uni.

Pour construire ce modèle, nous considérons une fonction croissante h qui effectue un changement d'échelle non uniforme de l'axe temporel : contraction peu importante du temps au début de la série, contraction du temps importante vers la fin de la série. Dans la pratique, une fonction logarithmique $h(t) = \log(t+1)$ sera utilisée afin d'obtenir les résolutions d'analyse souhaitées. La construction du modèle est alors simple : estimer un modèle à noyaux non pas sur les données initiales, les $(t_i^j)_{1 \leq i \leq k_j}$, mais sur les données transformées $(h(t_i^j))_i$, puis reprojeter cette estimation dans l'espace initial.

Nous notons $f^j(t)$ la fonction densité de la variable aléatoire T^j. Pour produire une estimation de cette densité, nous nous intéressons à la distribution de la variable aléatoire $h(T_i^j)$, que nous estimons grâce à un modèle à noyaux classique $\hat{n}^j(.)$. Une estimation de la densité $f^j(t)$ est alors donnée par (Charpentier et Flachaire, 2014) :

$$\hat{f}^j(t) = \frac{\hat{n}^j(h(t))}{|(h^{-1})'(h(t))|} = \hat{n}^j(h(t))h'(t)$$

où h^{-1} est la fonction réciproque de h, et h' sa dérivée. Pour $h(t) = \log(t+1)$, on obtient donc $\hat{f}^j(t) = \hat{n}^j(\log(t+1))/(t+1)$.

On peut noter qu'une approche relativement équivalente, mais ne nécessitant pas d'estimation dans un espace transformé est envisageable. Il est en effet possible de construire un modèle à noyaux dans l'espace initial avec des espacements entre noyaux et des largeurs de noyaux variables. Cette approche nécessite cependant un développement informatique spécifique, alors que l'approche basée sur la transformation des données par h permet de s'appuyer sur des bibliothèques informatiques éprouvées.

Dans la pratique, nous souhaitons obtenir une représentation vectorielle de cette distribution. Nous discrétisons chaque $\hat{f}^j(t)$ sur l'intervalle $[0, t_{sup}]$ en 25 points (t_{sup} peut être choisi comme la valeur du dernier tweet ou par un quantile). Pour obtenir les points de discrétisation de $\hat{f}^j(t)$, l'intervalle $[0, h(t_{sup})]$ est tout d'abord uniformément discrétisé en 25 points, puis chacun de ces points est reprojeté dans l'espace initial par la transformation réciproque h^{-1}.

Pour le paramètre de lissage (largeur des noyaux dans l'espace transformé), nous avons choisi une valeur unique pour tous les modèles à noyaux. La régle de Silverman (voir (Silverman, 1986), équation (3.31), page 48) a été utilisée pour déterminer une largeur adaptée pour chaque modèle, puis une valeur proche de la médiane a finalement été sélectionnée.

Sur la figure 2, on retrouve pour quatre articles la densité estimée par le modèle. En abscisse, la valeur des 25 points de discrétisation est indiquée (la première valeur n'est pas reportée). Les données sont décrites avec une résolution de l'ordre de la minute pour les premières variables, et de l'ordre du mois pour les dernières (56 jours séparent les deux derniers points de discrétisation).

Chaque point rouge au dessous des courbes représente un tweet. Une échelle logarithmique est utilisée pour représenter les données sur l'axe des abscisses. De fait, sur cette visualisation,

une répartition de tweets peut paraître uniforme, mais il n'en est rien : deux graduations successives ne représentent pas un même intervalle de temps en fonction de leur distance à l'origine.

3 Classification non supervisée

Après la discrétisation des modèles à noyaux, nous obtenons un tableau de données de 6 177 individus (les articles) et de 25 variables (les intervalles de temps). Nous avons appliqué une classification ascendante hiérarchique sur les articles afin d'identifier des comportements types de circulation d'événements en réponse à la publication d'un article. La distance euclidienne ainsi que le critère de Ward ont été utilisés. À partir de cette classification hiérarchique, une partition à 12 classes a été extraite. Le choix de 12 classes a permis d'obtenir une bonne homogénéïté dans chaque classe.

La figure 1 permet de juger de la qualité de représentation des données, et d'analyser les résultats de la classification. Sur cette *heatmap*, les valeurs jaunes correspondent à des valeurs élevées de la densité, alors que les valeurs rouges correspondent à des valeurs proches de 0. On distingue des différences de comportements à différentes échelles. Tout d'abord, des articles pour lesquels l'activité Twitter est concentrée sur les premiers instants suivant la publication de l'article (approximativement durant la première heure). Ensuite, les articles pour lesquels l'activité s'étend sur plusieurs heures, voire même sur plusieurs jours.

Afin de mieux décrire les différents comportements, nous proposons les *heatmaps* restreintes à certaines classes. La figure 3 décrit les classes 9, 3, 7 et 1 dont les centroïdes sont représentés par la figure 2. Pour ces classes, on peut extraire les caractérisations suivantes :

— le centroïde de la classe 9 montre un pic unique. L'échelle logarithmique sur l'axe des abscisses masque l'aspect asymétrique de la distribution. Cette classe se caractérise par une majorité de tweets publiés très tôt après la publication de l'article associé : la moyenne des médianes des articles se situe à 1h. La dispersion est très faible : l'écart interquartile moyen est de 2h ;
— le centroïde de la classe 3 montre une dispersion plus importante que celui de la classe 9. Cette classe se caractérise par une arrivée un peu plus tardive des tweets : la médiane moyenne est à 6h. La dispersion est relativement importante : l'écart interquartile moyen est de 20h ;
— le centroïde de la classe 7 montre l'extrême sensibilité aux occurrences en début de distribution : un seul tweet provoque un pic malgré le lissage du modèle à noyaux. Après ce pic initial, l'activité Twitter lié à l'article se maintient plusieurs heures. La classe 7 se caractérise par une majorité de tweets publiés tardivement : la médiane moyenne se situe à 11h. La dispersion est importante : l'écart interquartile moyen est de 27h ;
— la classe 1 se caractérise par de nombreux articles publiés vers minuit (au moins un quart). Les premiers tweets sont publiés plusieurs heures plus tard : en moyenne presque 6h après la publication de l'article. La médiane moyenne se situe à 33h, valeur beaucoup plus importante que pour les autres classes, ce qui indique une activité très décalée dans le temps. La dispersion est très importante : l'écart interquartile moyen est de 42h.

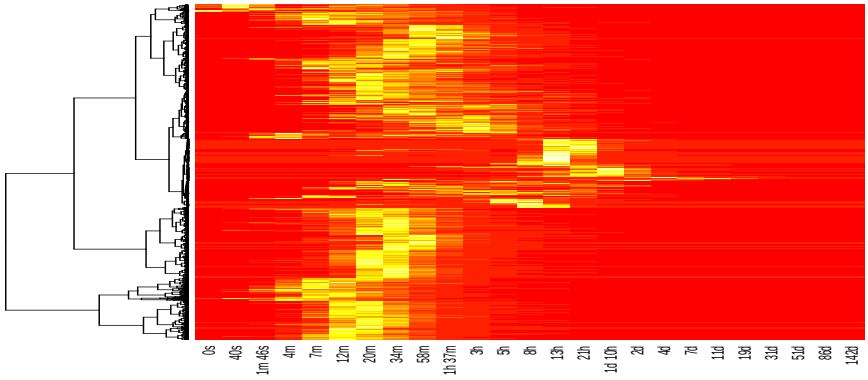


FIG. 1 – *Heatmap de la classification (la couleur jaune correspond à des valeurs élevées)*

4 Conclusion

Les premiers résultats de ce travail exploratoire sont très encourageants. Il reste notamment un travail d'interprétation et de description des classes en intégrant les méta-données associées aux tweets et aux articles. En effet, pour ce travail préliminaire, nous avons uniquement utilisé l'horodatage de publication des articles et des tweets, il serait intéressant de s'appuyer par exemple sur la thématique de l'article ou de différencier les tweets des retweets.

D'un point de vue technique, nous avons testé une représentation des données Twitter basée sur une échelle logarithmique. Celle-ci permet d'adapter la résolution d'analyse de façon à privilégier un niveau de détail plus élevé pour les tweets proches de la publication de l'article. Il nous reste à investiguer l'influence de trois paramètres : le paramètre de lissage des noyaux, la dimension des vecteurs de description, et la métrique utilisée. Actuellement, la classification obtenue utilise une métrique euclidienne, nous envisageons de tester par la suite d'autres dissimilarités, et notamment DTW très adaptée à la comparaison de séries temporelles.

Références

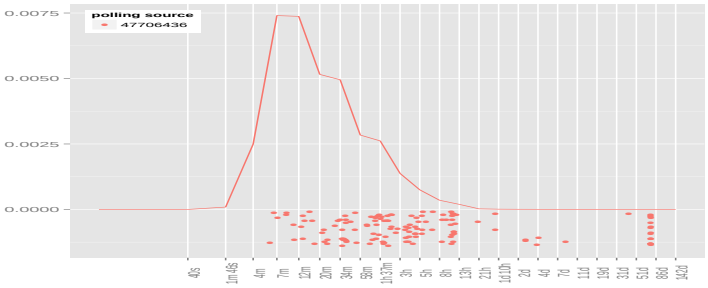
Chae, J., D. Thom, Y. Jang, S. Kim, T. Ertl, et D. S. Ebert (2014). Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics 38*, 51–60.

Charpentier, A. et E. Flachaire (2014). Log-Transform Kernel Density Estimation of Income Distribution. working paper or preprint.

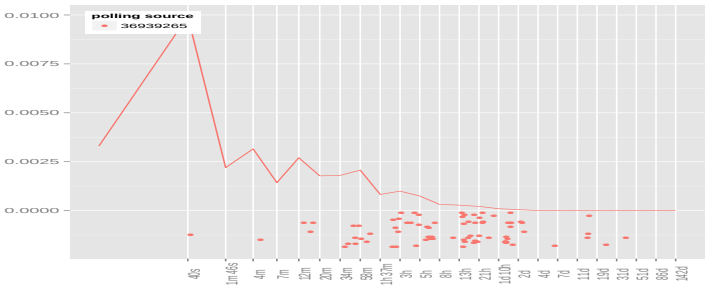
Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London : Chapman & Hall.

classe 9

classe 3



classe 7



classe 1

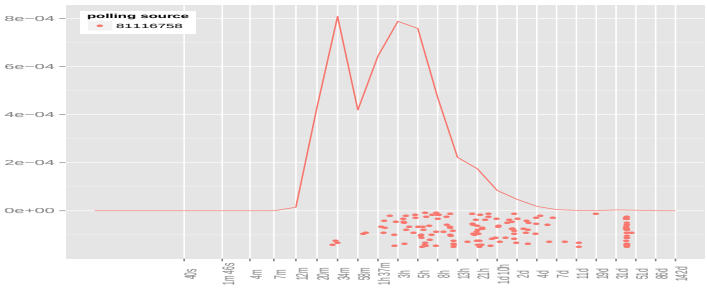


FIG. 2 – *Centroïdes des classes 9, 3, 7 et 1*

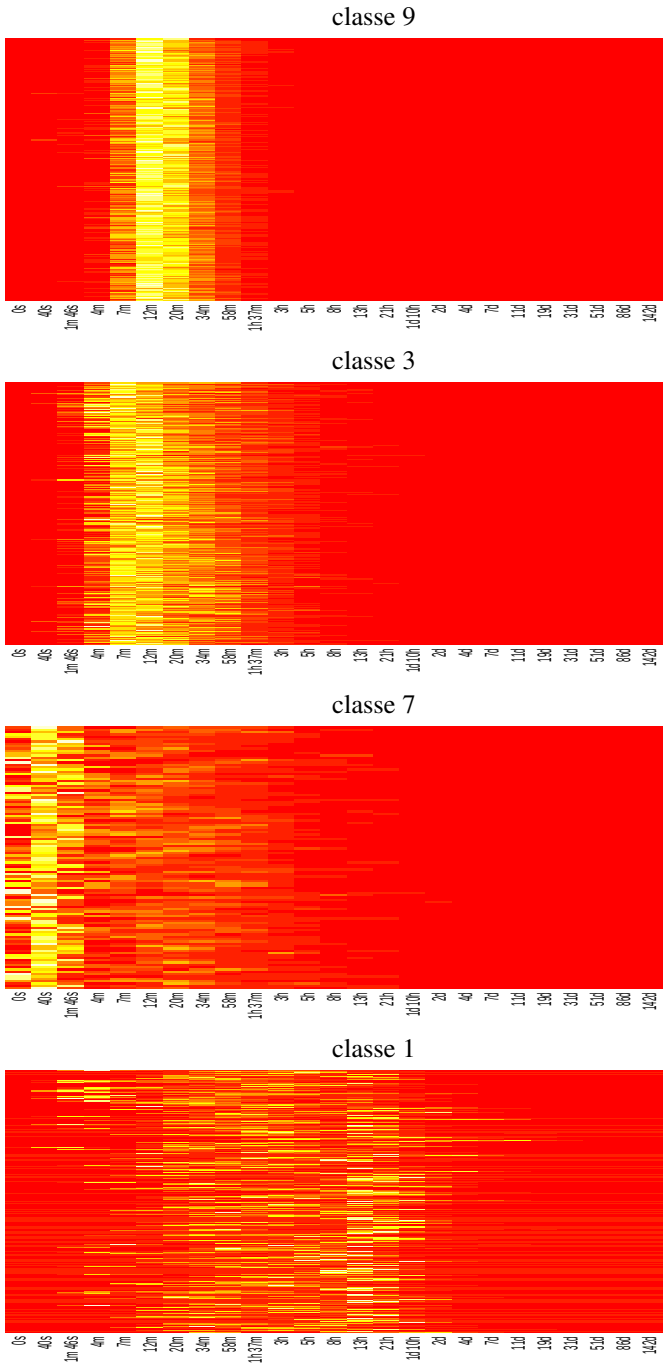


FIG. 3 – *Heatmap des classes 9, 3, 7 et 1*

Représentation de données temporelles par un modèle à échelle logarithmique

Summary

In this work, we are interested in tweets associated to a source event (online publication), and we seek to represent the temporal distribution of tweets thanks to a model with a variable resolution: high resolution at the beginning of the time serie, and coarse resolution at the end. For this purpose, we use a kernel model based on a logarithmic transformation of data.

Formation semi-automatique de classes sémantiques couvrantes pour enrichir une ontologie de domaine

Olena Orobinska[*,**] Jean-Hugues Chauchat[*] et Natalia Sharonova[**]

[*]Université de Lyon, ERIC, 5 av Pierre Mendès-France 69676, Bron Cedex, France
Olena.Orobinska@univ-lyon2.fr,
http://eric.ish-lyon.cnrs.fr/
[**]KhPI, 21 avenue Frunze, 61002, Kharkiv, Ukraine
sharonova@kpi.kharkov.ua
http://web.kpi.kharkov.ua/iks/

Résumé. Nous proposons une méthode hybride, semi-automatique, qui utilise l'intersection de classes sémantiques de noms et de verbes, classes construites sur le lexique du domaine, et une application qui complète un noyau d'ontologie à partir d'une liste de concepts initiaux et complète ce noyau par de nouvelles entités détectées dans un large corpus de textes de normes internationales de Sécurité Radiologique. Les résultats confirment le rôle important de la modélisation linguistique initiale ; ils montrent que les ressources lexicales externes, disponibles en ligne, peuvent contribuer efficacement à la résolution du problème de la désambiguïsation lexicale.

1 Introduction

1.1 L'enjeu technique et informatique

Après leur première apparition dans la communauté informatique au milieu des années 1980, les ontologies sont devenues aujourd'hui le cœur d'un grand nombre de systèmes d'information : Simperl et al. (2010). Pour Gruber *Une ontologie est la spécification formelle d'une conceptualisation* Gruber (1995). Aujourd'hui la description d'une spécification formelle est bien établie (la sémantique portée sur les formules syntaxiques du langage formel choisi), mais la définition des concepts associés à la sémantique d'un domaine, reste un enjeu ouvert dans le domaine de l'ingénierie des ontologies notamment dans l'apprentissage des ontologies.

Le cycle de vie d'une ontologie comprend plusieurs étapes ; chacune d'elles a ses propres objectifs et garantit le succès des suivantes. Mais l'importance de la conceptualisation d'un domaine (association univoque des libellés aux catégories des objets, abstraits ou réels, du domaine), est soulignée par de nombreux auteurs : Aussenac-Gilles (2005), Cimiano et al. (2014).

Les aspects méthodologiques généraux sont étudiés dans toute une série de travaux : García-Silva et al. (2014) ; Nicola et al. (2009) ; Simperl et al. (2010) . Une des tendances est l'unification de tous les processus impliqués dans la construction d'une ontologie afin d'économiser

les temps de conception et de calcul. Ainsi Hailpern et Tarr (2006) proposent MDD (model-driven development), une stratégie basée sur des modèles pour la prise des décisions gestionnaires et techniques de développement des systèmes d'information (SI). Le paradigme actuel de programmation est orienté sur le développement de SI munis, a priori, de connaissances sur l'objet à traiter ou d'un modèle d'objet ; l'ontologie en est une des formes. Dans l'élaboration du modèle, le travail des experts est fondamental mais couteux. Notre contribution rejoint une autre tendance de développement des SI : on cherche à remplacer le travail des experts, là où c'est possible, par une exploitation directe des ressources numériques en utilisant des méthodes d'apprentissage, nommées ici méthodes d'apprentissage des ontologies (OL - ontology learning).

1.2 L'enjeu du domaine

Le domaine de la sécurité radiologique a pour objectif la protection de la santé de la population, et du personnel qui travaille avec des sources radioactives, contre les impacts néfastes des rayonnements ionisant ; il édicte des principes et des normes : Casanova (2008). La gestion des risques radiologiques, réglementée par la série des normes de sureté de l'International Atomic Energy Agency (IAEA) et d'ISO, doit faciliter la prise de décision ; elle suppose la construction et la maintenance d'une base de connaissances fondée sur une ontologie du domaine (ou sur plusieurs ontologies de sous-domaines spécifiques).

La section 2 introduit l'état de l'art sur les techniques d'extraction de termes et de relations à partir de ressources textuelles ; nous présentons notre approche sur la construction du noyau d'ontologie et son extension avec des nouvelles instances dans la section 3 ; la section 4 détaille et illustre l'implémentation de notre méthode et la section 5 souligne des problèmes à résoudre et introduit quelques perspectives.

2 État de l'art

Nous nous limitons ici aux méthodes d'apprentissage d'une ontologie de domaine fondées sur la terminologie de ce domaine.

Les termes extraits du corpus peuvent être utilisés pour dénommer les concepts de l'ontologie et l'instancier. Traditionnellement l'apprentissage d'ontologie (OL - ontology learning) a utilisé des méthodes statistiques. Ensuite les méthodes linguistiques (syntaxiques) ont été largement appliquées en raison de la performance croissante des analyseurs morphologiques et syntaxiques, par exemple TreeTagger : Buitelaar et Cimiano (2008), Brill (1992), Schmid (1994).

Au cours des dernières années, de plus en plus de ressources lexicales lisibles par les ordinateurs (machine readable) sont devenues disponibles : les bases de connaissances génériques (OpenCyc1), les thésaurus génériques ou spécialisés (UMLS2), les dictionnaires de synonymes (WordNet3, CRISCO, etc.), les dictionnaires multilingues (BabelNet4) etc. En retour, ces outils ont stimulé l'émergence de nouvelles techniques lexicales.

Les méthodes d'apprentissage visent soit à créer une ontologie sans a priori (from scratch), soit à instancier une ontologie de base, étape par étape, à l'aide d'experts du domaine. Jusqu'à récemment on pouvait citer deux stratégies principales pour construire une ontologie. La première est dite stratégie *Layer cake*, Cimiano et al. (2005) où chaque étape utilise les résultats

des tâches précédentes. On commence par l'extraction des termes-candidats qui, ensuite, se groupent et forment la hiérarchie des relations. Les règles d'apprentissage sont aussi dérivées de ces données initiales (les termes-candidats). Les supports les plus réussis, construits sur cette approche et qui continuent de se développer, sont KAON, TEXT2ONTO : Bozsak et al. (2002) et Haase et Motik (2005). Une autre approche met en œuvre une phase de conceptualisation qui passe par modélisation linguistique des ressources disponibles pour découvrir dans le corpus les entités de concepts à travers des relations prédéfinies. Cette approche, ainsi que les outils de mise en œuvre tels que TERMINAE et CHAMELEON, sont présentés dans Mondary et al. (2008) et Aussenac-Gilles et al. (2000). Un ouvrage tout récent décrit les progrès et la maturation actuelle de cette vision ; les auteurs déploient les principes généraux des systèmes de traitement des textes en langue naturelle ; l'ontologie linguistique est au cœur du système ; les auteurs ont élaboré le modèle Lemon qui applique ces idées : Cimiano et al. (2014).

Toutefois la classification des relations sémantiques n'est pas encore entièrement résolue car toute relation peut être exprimée linguistiquement de façons différentes. Certaines relations sont actuellement acceptées : synonymie, relations hiérarchiques ou taxonomiques telles que hyperonymie, méronymie, etc.

De nombreux ouvrages proposent leurs propres classements des relations sémantiques. Ainsi Rosario et Hearst (2001) proposent 38 classes qui décrivent les relations sémantiques possibles dans un groupe nominal entre le nom central et ses dérivés. Moldovan et Novischi (2004) proposent le système de 35 classes pour classer les relations sémantiques dans les groupes nominaux ; plus tard ce schéma a été repris et développé par Girju et al. (2005). Pour caractériser les relations sémantiques entre les verbes, Pantel et Ravichandran (2004) introduisent cinq classes.

Parmi les algorithmes visant à déceler les relations sémantiques mentionnons :

- l'analyse distributive (DA, Distributional Analysis) qui permet de mettre les termes en relation, avec les types de relations différentes : Morlane-Hondère et Fabre (2012) ; on est alors contraint de faire la distinction entre les relations syntagmatique (comme les collocations où les relations nom-verbe) et les relations paradigmatiques (comme la synonymie, hyperonymie) : Fabre et Bourigault (2006) ;

- l'exploitation de patrons lexico-syntaxiques qui permettent d'avoir une bonne précision. La contrainte principale est qu'ils sont limités par les contextes où ils peuvent être utilisés : Orobinska et al. (2013). Habituellement, un seul type de relation peut être déterminé avec cette méthode : Perinet et Hamon (2013) .

Dans les dernières années, beaucoup de chercheurs se sont penchés sur le repérage et l'exploitation des relations non taxonomiques en vue d'installer ou d'enrichir une ontologie. Par exemple Makki et al. (2009) proposent une méthode semi-automatique de peuplement d'une ontologie, basée sur l'application de règles de reconnaissance des instances où les relations sémantiques entre les concepts sont prédéfinis par des experts. Dans Kavalec et al. (2004) la méthode d'extraction des relations caractéristiques entre les concepts se base sur un indicateur statistique qui mesure le niveau d'importance de chaque verbe afin de caractériser le lien entre les concepts. A ces mêmes fins, Punuru et Chen (2012) proposent d'exploiter les triplets du type (sujet, prédicat, objet) où la signification des verbes est estimée par le rapport de vraisemblance.

Plusieurs études montrent l'efficacité de l'utilisation des dictionnaires électroniques pour l'estimation de la similarité des instances-candidats et la résolution du problème d'ambiguïté

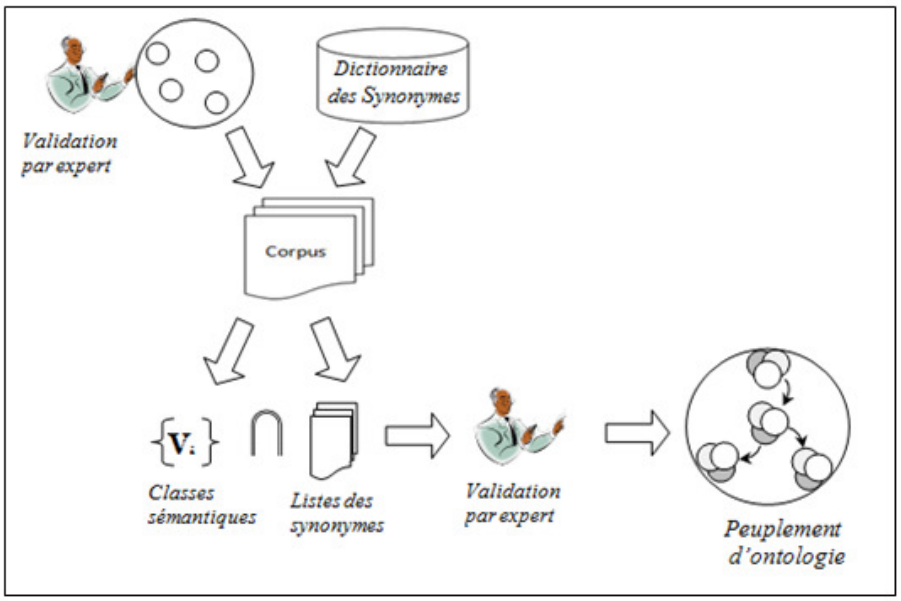


FIG. 1 – *Le schéma de la méthode.*

des mots. Par exemple, Flati et Navigli (2013) et Velardi et al. (2013)proposent d'utiliser largement BabelNet, pour les différents aspects de traitement linguistique. En particulier, ils proposent une approche unifiée multilingue, basée sur un graphe, qui procure le liage des entités et la désambiguïsation des sens de mots basée sur une identification floue de significations, combiné à l'heuristique la plus dense d'un sous-graphe. L'approche permet de sélectionner des interprétations sémantiques de haute cohérence : Moro et al. (2014). Fabre et Bourigault (2006) présentent les méthodes utilisant WordNet pour repérer la synonymie entre les mots.

3 Description de l'approche proposée

Nous proposons ici une méthode permettant de récupérer dans le corpus les éléments les plus pertinents pour peupler l'ontologie de domaine, en résolvant le problème de l'ambigüité des mots (WSD). Notre solution intègre les avantages des différentes approches pour gagner en généralité, fiabilité et simplicité.

Dans cette section nous présentons l'algorithme proposé figure 1. Chaque étape est détaillée dans les sections suivantes. L'expert n'intervient qu'aux première et dernière étapes ; entre les deux, la démarche est entièrement automatique.

Le processus (présenté figure 1) suit les étapes suivantes :

1. La constitution et l'étiquetage de corpus, le choix du dictionnaire généraliste de synonymes, ainsi que du glossaire du domaine.

2. La constitution du noyau de l'ontologie : à l'issu de cette étape on installe les liens entre les concepts par les verbes caractéristiques associés. Pour cela, on réalise une analyse syntaxique superficielle, basée sur les résultats d'un parseur morphologique. Nous n'utilisons pas d'analyseur syntaxique car les résultats sont décevants, les phrases du corpus étant souvent trop longues et complexes. On cherche dans le corpus les triplets (sujet, prédicat, objet), où sujet et objet contiennent les mots présentant les concepts génériques. Pour nous, ici, un sujet est tout le groupe nominal qui se trouve à gauche d'un verbe dans le fragment de texte limité par deux signes de ponctuation (hormis la virgule) ; de même, un objet est constitué de tout le groupe nominal qui se trouve à droite d'un verbe. Par exemple dans la phrase : *Les estimations globales du risque stochastique restent encore essentiellement les mêmes* le sujet du verbe *rester* est la locution *les estimations globales du risque stochastique* ; et dans la phrase : *La Commission avait quantifié le risque des effets stochastiques des rayonnements et proposé un système de limitation de dose avec ses trois principes de justification* l'objet du verbe *quantifier* est la locution *le risque des effets stochastiques des rayonnements*.

3. Construction des patrons relationnels de type SVO. Dans le domaine technique développé (la sécurité radiologique), le lexique comprend des termes composés de plusieurs mots, par exemple : *dose absorbée individuelle* , *brûlure radiologique étendue* , *destruction due aux rayonnements* , etc. C'est pourquoi on applique classiquement les patrons morphologiques à les groupes nominaux de type *NOM ADJ* ou *NOM ADJ PPP* etc., où *NOM* , *ADJ* et *PPP* sont les marqueurs étiquetant les mots du corpus à l'aide d'un parseur (dans notre exemple, ils correspondent au noms, adjectifs et participes passées). Mais l'application des seuls patrons morphologiques a une pertinence faible car ce sont des structures syntaxiques communes de la langue, et pas caractéristiques d'un domaine particulier. Les résultats s'améliorent si, dans les patrons, on utilise des concepts figurant dans la liste de l'expert, mais, même dans ce cas, la pertinence moyenne ne dépasse pas 50%. Pour améliorer les résultats, nous utilisons les verbes caractéristiques associés dans le corpus aux concepts de notre liste initiale. Pour éviter la perte du rappel sans diminuer la précision des résultats nous créons des classes de verbes proches sémantiquement (ou classes de verbes-synonymes) ; exemple, la liste des verbes du corpus qui expriment la causalité (*causer, provoquer, déclencher, engendrer, susciter*). La combinaison des patrons morphosyntaxiques et les classes des verbes, permet d'étendre la liste enrichie des concepts en choisissant dans le corpus de nouveaux candidats-termes avec une bonne pertinence. Ceci est illustré et détaillé aux §4.2 et 4.3.

4. La pré-validation des concept-candidats se fait automatiquement : on part d'une classe de verbes et on récupère dans le corpus tous les noms associés aux verbes de cette classe. De la liste des noms dégagés on ne sauvegarde que ceux qui sont les synonymes du concept associé avec la classe donnée des verbes. Exemple : le concept *dommage* est associé à la classe des verbes causatives (*causer, provoquer, déclencher, engendrer, susciter*). Dans le corpus, une longue liste de mots est associée à ces verbes (*distorsion, défaillance, développement, événement*, etc.). On retient les noms *distorsion* et *défaillance* car ils sont les synonymes du concept *dommage* et on élimine les autres. L'expert valide finalement les résultats proposés par l'algorithme.

4 Implémentation et résultats

4.1 Étape préparatoire : un corpus étiqueté

Nous avons constitué un corpus original de textes français sur le domaine de la Sécurité Radiologique ; il réunit des normes de sûreté et des rapports de l'*IAEA* et de la *Commission Internationale de Protection Radiologique*, ainsi que des articles de revues spécialisées publiés depuis 1999. Notre corpus contient un million et demi de mots.

Ensuite, les textes ont été étiquetés avec TreeTagger Schmid (1994) de façon à obtenir des fichiers sous la forme *mot/POS/lemme mot/POS/lemme* etc. où *mot* correspond à tout mot du texte, *lemme* présente sa forme normale (l'infinitif pour les verbes ; le singulier pour les noms, la forme masculin-singulier pour les adjectifs, etc.) La ponctuation est également enregistrée pour pouvoir distinguer les phrases. Le choix en faveur de TreeTagger a été fait après avoir testé plusieurs outils tel que Stanford et Brill, et à partir des résultats présentés dans Falk et al. sur les 7 parseurs pour le français ; la qualité de TreeTagger est particulièrement bonne pour les noms et les verbes c'est à dire pour les parties du discours qui nous intéressent en priorité.

Par exemple, pour le texte initial : *détriment causé au personnel médical est pris en compte de façon subsidiaire*. On obtient la structure étiquetée : *détriment/NOM/détriment causé/PPP/ causer au/PRP/au personnel/NOM/personnel médical/ADJ/médical est/VER/être pris/PPP/ prendre en/PRP/en NOM/compte de/PRP/de façon/NOM/façon subsidiaire/ADJ/ subsidiaire/ SENT/.*

4.2 Constitution du noyau de l'ontologie

L'installation du noyau d'ontologie suppose l'énumération des concepts génériques et des prédicats dans un des formats descriptifs tels qu'OWL. Comme, généralement, les concepts supérieurs sont peu nombreux Velardi et al. (2013) et qu'ils conditionnent la qualité des résultats, nous préconisons d'installer le modèle initial en coopération avec l'expert du domaine. Notre liste initiale contient neuf éléments, LC=(*dommage, exposition, contrôle, personnel, protection, rayonnement, sûreté, source, risque*).

Ensuite on réalise une analyse syntaxique rapide du corpus, basée sur les résultats du traitement par le parseur morphologique : on cherche les triplets (sujet, prédicat, objet) des phrases où le sujet et l'objet contiennent les concepts génériques. Les prédicats sont retenus si le poids du verbe dépasse un seuil choisi empiriquement. Le poids donné à chaque verbe est calculé par un coefficient combinant l'information mutuelle (1) et l'indice de Jaccard (2) :

$$IM(c_i v_j) = \sum f_{ij} * log \frac{f_{ij}}{f_{c_i} * f_{v_j}} \tag{1}$$

$$k_{ij} = \frac{f_{ij}}{N_i + N_j} \tag{2}$$

Ici c_i et v_j correspondent à un concept et un verbe respectivement ; f_{ij} est la fréquence de cooccurrence du concept c_i avec le verbe v_j ; f_{c_j} et f_{v_j} sont les fréquences propres du concept c_i et du verbe v_j ; N_i et N_j correspondent au nombre des occurrences dans le corpus du concept c_i et du verbe v_j respectivement.

Plusieurs résultats illustratifs sont présentés dans la table 1.

Paires sujet-objet	Prédicat liant (verbe)
rayonnement - dommage	indure, provoquer, conduire, diagnostiquer
rayonnement - exposition	entraîner, maintenir, risquer
protection - exposition	réduire, contrôler
source - protection	garantir
risque - exposition	entraîner

TAB. 1 – *Exemples de prédicats liant des concepts de noyau d'ontologie.*

4.3 Création des classes de verbes et extension du noyau d'ontologie

Pour les concepts et les verbes sélectionnés à l'étape précédente, on forme les groupes de synonymes présents dans le lexique de corpus. Pour cela nous avons utilisé le Dictionnaire Électronique des Synonymes.

Comme la relation de synonymie dépend de contexte nous avons conservé uniquement les synonymes réciproques au sens où deux mots a et b sont *synonymes réciproques* si, dans le dictionnaire, a figure parmi la liste des synonymes de b, ET b figure dans la liste des synonymes de a.

Les tables 2 et 3 présentent les résultats pour des verbes et des noms.

Partie du discours	Nb des unités Étiquetés	Étiquetés correctement	Présent dans le dictionnaire CRISCO	Manquant dans le dictionnaire CRISCO
Verbes	1093	1029	1009	20
Noms	2400	2074	2049	25

TAB. 2 – *Etiquetage des noms et verbes du corpus, et présence-absence dans CRISCO.*

	Total des mots	Nombre total des synonymes réciproques dans CRISCO					
		0	1	2-4	5-12	13-20	>20
Verbes	1009	42	46	137	318	213	253
Noms	2040	141	205	408	648	325	313

TAB. 3 – *Nombre de synonymes réciproques selon la taille des groupes de verbes et de noms.*

Pour installer les classes sémantiques des verbes nous avons calculé le taux de similarité intra-classe dans notre corpus (la similarité dépend du contexte).

Le niveau de similarité entre les verbes est calculé avec la formule de cosinus (3). Ici les listes des synonymes pour chaque mot-entré sont présentées par les vecteurs (le mot-entré est le mot avec lequel on interroge le dictionnaire des synonymes).

$$SimCos(v_l, v_i) = \frac{||V_l \bigcap V_i||}{\sqrt{\|V_l\| \times \|V_i\|}} \quad (3)$$

où $\| \|V_l\| \|$ est la cardinalité du vecteur du mot-entré l ; $\| \|V_i\| \|$ est la cardinalité de vecteur de son i-ème synonyme ; $\| \|V_l\| \|$ ET $\| \|V_i\| \|$ est le résultat de l'intersection (nombre de synonymes réciproques).

Une classe est établie pour les verbes dont le taux de similarité α_{ij} est supérieur à 0.25 ; Le seuil été défini expérimentalement.

Exemple. Le cosinus de similarité entre le verbe *protéger* et le verbe *préserver* est calculé comme suit : Le dictionnaire CRISCO donne 27 synonymes pour *protéger* et 17 pour *préserver* ; 13 verbes figurent à la fois dans ces deux listes ; le cosinus de similarité est donc égal à 0,61.

La règle d'inférence est formulée comme suit : les synonymes d'un concept associés dans le corpus avec la même classe des verbes sont suggérés comme candidats pour libeller le concept et enrichir le noyau de l'ontologie. L'expert intervient enfin pour valider les nouveaux termes.

Des exemples sont présentés dans la table 4.

Concept	Classe sémantique du prédicat	Nouveaux noms
dommage	causer, provoquer, déclencher, engendrer, susciter	distorsion, défaillance
protection	réduire, diminuer	préservation, surveillance
sécurité	contrôler, examiner, inspecter	assurance, confiance
rayonnement	limiter, réduire, déterminer	radiation, irradiation

TAB. 4 – *Exemples de résultats de l'instanciation du noyau de l'ontologie.*

Il est à noter que les synonymes de chaque concept que nous découvrons ne forment pas une hiérarchie dans l'ontologie ; nous découvrons des labels différents pour nommer chacun de ces concepts initiaux. La précision des résultats est supérieure à 90%, néanmoins le rappel moyen est de l'ordre de 50%.

5 Conclusion et travaux futurs

Nous avons créé un grand corpus sur un domaine technique puis proposé et mis en œuvre une méthode semi-automatique pour construire une ontologie de domaine avec peu de connaissances au départ :

1. détecter des catégories des relations importantes liant les concepts du domaine ;
2. créer les classes sémantiques de verbes et de concepts pour le noyau d'ontologie ;
3. créer les patrons morphosyntaxique (PMS) ;
4. en utilisant ces PMS, créer les règles des reconnaissances pour l'extraction de termes et l'enrichissement de l'ontologie.

La nouveauté principale de la méthode consiste dans le changement de l'ordre traditionnel des étapes : nous installons d'abord les structures relationnelles entre les concepts, puis celles-ci permettent de trouver les instances, et non l'inverse.

Les avantages de la méthode sont :

- Elle est cohérente et permet de construire l'ontologie à partir de zéro, avec un minimum de connaissances préalable.

- Elle donne des résultats d'une haute précision.

- Le travail des experts est réduit à l'initialisation puis à la validation des résultats finaux.

Parmi les inconvénients de la méthode on peut noter les difficultés d'analyse syntaxique des phrases longues (et les analyseurs connus, tels que Sanford ne donnent pas de meilleurs résultats). Par ailleurs, le rappel est relativement peu élevé. Ce sont les problèmes à résoudre dans les travaux futurs.

Références

Aussenac-Gilles, N. (2005). *Méthodes ascendantes pour l'ingénierie des connaissances*. Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France.

Aussenac-Gilles, N., B. Biébow, et S. Szulman (2000). Revisiting ontology design : a method based on corpus analysis. In *12th International Conference on Knowledge Engineering and Knowledge Management, Juans-Les-Pins (F), 03/10/2000-06/10/2000*, pp. 172–188. Springer Verlag.

Bozsak, E., M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz, et V. Zacharias (2002). Kaon - towards a large scale semantic web. In K. Bauknecht, A. M. Tjoa, et G. Quirchmayr (Eds.), *E-Commerce and Web Technologies, Third International Conference, EC-Web 2002, Proceedings*, Volume 2455 of *LNCS*, Berlin, pp. 304–313. Springer.

Brill, E. (1992). A simple rule-based part of speech tagger. In *In Proceedings of the Third Conference on Applied Natural Language Processing*.

Buitelaar, P. et P. Cimiano (2008). *Ontology Learning and Population : Bridging the Gap Between Text and Knowledge*, Volume 167. Ios Press Inc.

Casanova (2008). Impact de la norme ISO 20553 sur le suivi de l'exposition interne. Technical report, Société Française de Radio-Protection - Section de Protection Technique, Union Internationale des Chemins de Fer - Paris.

Cimiano, P., A. Hotho, et S. Staab (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Int. Res. 24*(1), 305–339.

Cimiano, P., C. Unger, et J. P. McCrae (2014). *Ontology-Based Interpretation of Natural Language*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Fabre, C. et D. Bourigault (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *Actes de la 13e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2006)*, pp. 121–129.

Falk, I., D. Bernhard, C. Gérard, et R. Potier-Ferry. Étiquetage morpho-syntaxique pour des mots nouveaux. In *Actes de la 21 Conférence de Traitement Automatique des Langues Naturelle, TALN-2014), Marseille, France, year = 2008*.

Flati, T. et R. Navigli (2013). Spred : Large-scale harvesting of semantic predicates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1 : Long Papers*, pp. 1222–1232.

García-Silva, A., L. J. García-Castro, A. G. Castro, et Ó. Corcho (2014). Social tags and linked data for ontology development : A case study in the financial domain. In *4th International Conference on Web Intelligence, Mining and Semantics (WIMS 14), WIMS '14, Thessaloniki, Greece, June 2-4, 2014*, pp. 32 :1–32 :10.

Girju, R., D. I. Moldovan, M. Tatu, et D. Antohe (2005). On the semantics of noun compounds. *Computer Speech & Language 19*(4), 479–496.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud. 43*(5-6), 907–928.

Haase, P. et B. Motik (2005). A mapping system for the integration of owl-dl ontologies. In *Proceedings of the First International Workshop on Interoperability of Heterogeneous Information Systems*, IHIS '05, New York, NY, USA, pp. 9–16. ACM.

Hailpern, B. et P. Tarr (2006). Model-driven development : The good, the bad, and the ugly. *IBM Syst. J. 45*(3), 451–461.

Kavalec, M., A. Maedche, et V. Svatek (2004). Discovery of lexical entries for non-taxonomic relations in ontology learning. In P. van Emde Boas, J. Pokorn, M. Bielikov, et J. Stuller (Eds.), *SOFSEM*, Volume 2932 of *Lecture Notes in Computer Science*, pp. 249–256. Springer.

Makki, J., A.-M. Alquier, et V. Prince (2009). Ontology population via NLP techniques in risk management. *International Journal of Humanities and Social Sciences 3*(3), 212–217.

Moldovan, D. I. et A. Novischi (2004). Word sense disambiguation of wordnet glosses. *Computer Speech & Language 18*(3), 301–317.

Mondary, T., S. Després, A. Nazarenko, et S. Szulman (2008). Construction d'ontologies à partir de textes : la phase de conceptualisation. In *19èmes Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, Nancy, France, pp. 87–98.

Morlane-Hondère, F. et C. Fabre (2012). Étude des manifestations de la relation de méronymie dans une ressource distributionnelle. In *Actes de la Conference de Traitement Automatique des Langues Naturelle, TALN, Grenoble, France*, pp. 169–182.

Moro, A., A. Raganato, et R. Navigli (2014). Entity linking meets word sense disambiguation : a unified approach. *TACL 2*, 231–244.

Nicola, A. D., M. Missikoff, et R. Navigli (2009). A software engineering approach to ontology building.

Orobinska, O., J.-H. Chauchat, et N. Charonova (2013). Enrichissement d'une ontologie de domaine par extension des relations taxonomiques à partir de corpus spécialisé. In G. A. de Cea et N. Aussenac-Gilles (Eds.), *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence TIA 2013*, pp. 129–137.

Pantel, P. et D. Ravichandran (2004). Automatically labeling semantic classes. In *HLT-NAACL*, pp. 321–328.

Perinet, A. et T. Hamon (2013). Hybrid acquisition of semantic relations based on context normalization in distributional analysis. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*, pp. 113–122.
English

Punuru, J. et J. Chen (2012). Learning non-taxonomical semantic relations from domain texts. *Journal of Intelligent Information Systems 38*(1), 191–207.

Rosario, B. et M. Hearst (2001). Classifying the semantic relations in noun compounds. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees.

Simperl, E. P. B., M. Mochol, et T. Burger (2010). Achieving maturity : the state of practice in ontology engineering in 2009. *IJCSA 7*(1), 45–65.

Velardi, P., S. Faralli, et R. Navigli (2013). Ontolearn reloaded : A graph-based algorithm for taxonomy induction. *Computational Linguistics 39*(3), 665–707.

Summary

We propose a hybrid, semi-automatic method which uses the intersection of semantic classes of nouns and verbs; the classes are built from the domain lexicon. An application completes a core ontology from a list of initial concepts and complete this core by new entities detected in a large corpus of international standards on Radiological Safety texts. The results confirm the important role of initial linguistic modeling; they show that external lexical resources, available online, can be effective in solving the problem of word sense disambiguation.

Mise au point d'un agrégateur de tweets par apprentissage semi supervisé d'ontologies juridiques

Pierre Saurel*, Yann Girard*,**
Lambert Pepin***

* SND Sciences, Normes, Décision(FRE 3593) Maison de la recherche - Paris 4
28, rue Serpente - 75006 Paris
pierre.saurel@paris-sorbonne.fr,
http://snd-sorbonne.fr/
** MathisTemis consulting
27 rue du faubourg poissonniere - 75009 Paris
http://mathisthemis.fr/
*** Laboratoire d'informatique de Nantes Atlantique.
Equipe DUKe (UMR CNRS 6241)
2, rue de la Houssiniere - BP 92208
44322 NANTES CEDEX 3
lambert.pepin@gmail.com

Résumé. Cet article présente les directions méthodologiques d'un système d'agrégation de tweet en cours d'implémentation. L'objectif est de catégoriser ces tweets pour les rattacher à des concepts juridiques organisés sous forme de structures ontologiques afin de repérer, à l'intérieur d'un flux commun de Tweets, ceux qui peuvent être rattachés à un concept juridique pertinent pour pouvoir finalement les isoler ou les retweeter. La nature floue et polysémique des concepts juridiques nous a amenés à structurer l'information à l'intérieur d'un tweet, en isolant des tokens afin d'extraire différents traitements. Cette structure permet un apprentissage non supervisé en le confrontant à une structure de connaissance a priori issue de notre expertise juridique. On a ainsi trois types de traitements: Une recherche de pertinence localisée, une analyse non supervisée de distributions, des connaissances a priori organisées sous la forme de classes et un système supervisé permettant de réviser les relations entre ces classes

1 Introduction

L'invitation de twitter *" What's happening ? Compose new tweet !"* (Que se passe-t-il ? Ecris donc un nouveau tweet !) a fait de cette outils de communication sociale une source d'information sur l'actualité en général, et l'actualité professionnelle en particulier (Power et al. (2011)) . Le domaine du droit n'échappe à cette règle. Certains arrêts ont un impact de nature à modifier les règles du jeu dans leur domaine. Certaines affaires d'actualité et la façon dont elles sont traitées peuvent elles aussi indiquer un changement de Jurisprudence

à venir. Le droit est une matière politique au sens étymologique du terme (qui organise la vie de la cité) et se tenir au courant est le devoir d'un bon légiste. Au-delà de s'informer, être capable d'informer autrui est un talent appréciable pour se faire connaitre. Encore faut-il avoir eu connaissance de la bonne information au bon moment. L'outil que nous développons est un agrégateur de données pouvant avoir une pertinence dans le domaine juridique. Il est restreint à l'examen des fils francophones afin de filtrer ces tweets pertinents et de les relier à un concept juridique approprié. Ce travail est encore très récents, le corpus de travail est encore en cours de constitution. Nous présenterons ici des résultats mais les choix méthodologiques que nous avons fait afin d'ouvrir une discussion avec la communauté. Dans un souci de tester nos choix durant la progression, nous stockons de façons régulières une partie de la production francophone sur twitter. Nous avons constitué un premier corpus de test contenant environs 120k tweets (à noël, en augmentation régulière d'environs 3 à 8k par jours)

2 Problématique

Hormis son intérêt pratique, cet outil aborde des problématiques pluridisciplinaires. La première concerne la détection d'un signal faible dans un flux de données : les tweets pertinents ne représentent qu'une infime partie du signal tweeter francophone journalier. Nous n'avons pas de chiffres définitifs, mais sur notre premier corpus ils sont aux alentours de 0,17La problématique de la recherche d'un signal faible dans un flux de réseaux sociaux notamment twitter est centrale pour tout outil de fouille de données chargé d'effectuer une veille ou une caractérisation de phénomène. Ce signal est souvent faible en représentation mais porteur de sens aussi bien pour le contrôle d'un marché que celui de l'information (Harryson et al. (2014)). La seconde problématique est d'obtenir représentation des concepts du droit sous forme d'une ontologie de concepts juridiques. Cette tâche met en œuvre une réflexion en psychologie appliquée autant qu'une expertise juridique afin de pouvoir définir le périmètre sémantique de chacun de ses concepts. Expertise d'autant plus cruciales que les sources du droit, l'ensemble des textes qui vont décrire le champ d'application d'une qualification, sont variés : la loi, la jurisprudence, les débats parlementaires, décrets d'application, etc. Finalement, l'établissement de liens entre une ontologie de concepts et les thèmes évoqués dans les tweets constitue une troisième problématique de recherche qui implique de spécifier un algorithme qui saura aussi bien prendre en compte les sujets dans les tweets que de rassembler les concepts dans un modèle.

3 Construction d'une ontologie des concepts juridiques

Les concepts juridiques se rattachent à des textes qui les décrivent soit spécifiquement, soit parce qu'ils définissent le périmètre d'application auquel se rattache ce concept. Les textes peuvent venir de différentes sources : les codes, le bulletin officiel, les arrêts de tribunaux de dernières instances ou des cours européennes, les directives européennes, mais aussi les rapports parlementaires, des articles scientifiques ou encore des publications sur des sites de grands cabinets d'avocats. La recherche de la similitude entre des classes et des instances comme les tweets posent des questions évoquées par Pierre (2014)dans ses travaux sur l'altérité radicale. Il s'agit notamment de trouver une approche adaptée au problème posé par le 'flou

du droit' selon l'expression de Delmas-Marty (1986). Afin de définir le périmètre d'application de chaque concept, nous avons choisi de les représenter sous forme de classes composées d'un lexique de termes associés. Ces termes sont extraits des textes principaux définissant le périmètre d'application du concept et sont enrichis d'un pondération, initialement attribuée à la main, puis raffiné par entraînement du système. En plus d'un champ lexical, les classes contiennent également des sources et des hashtags qui sont reconnus via notre expertise comme étant utilisé dans des discussions ou des annonces portant sur le domaine décrit par la classe. Cette liste est destinée à faire l'objet de révisions régulières notamment pour les hashtags qui sont souvent lié à l'actualité. Notre système semi-supervisé qui repose sur l'ontologie constituée par notre expertise juridique (supervisé) apporte une connaissance a priori utilisée pour catégoriser les tweets retenus (de façon non supervisée).

3.1 Exemple de catégorie juridique : Le concept juridique d'hébergeur

La qualification d'hébergeur pour une structure implique un ensemble de responsabilité, notamment celle de conserver les données de connections de ces utilisateurs à disposition éventuelles des autorités (jurisprudence européenne, précisés dans une directive et transposée dans l'article 6 de la LCEN, Loi (2010)). Inversement ce statut permet de s'affranchir de responsabilité lié à la connaissance obligatoire du contenu hébergé. Pendant longtemps le périmètre de la qualification d'hébergeur est resté flou et malgré différents travaux (révision de la loi Informatique et liberté, la loi sur la confiance dans l'économie numérique en 2004, Ordonnance (2015) JO (2015)) des incertitudes demeurent. Contrairement à des concepts comme 'la constitution' ou 'la volonté', pour définir un lexique, il convient de s'intéresser à de nombreux textes différents comme le rapport Pierre Sirinelli (Sirinelli et al. (2014)), la directive européenne 2001/29/CE (CJUE (2015)), l'arrêt Google de la cour de justice de l'union européenne, et même l'actualité puisqu'un projet législatif est conduit par Fleur Pellerin au moment où nous écrivons avec pour objectif de préciser le statut d'hébergeur.

Loi confiance économie numérique	3	droit d'auteurs	3
Internet	1	directive 89/104	3
Réseaux Sociaux	1	article 14	3
Plateforme Vidéos	2	directive 2000/31	3
Contenus	2	connaissance donnée	3
Liberté d'expression	2	contrôle donnée	3
Hébergeurs	3	rôle actif	3
Editeurs	2	arrêt 12 juillet 2011	3
Consommateurs	2	arrêt 12/06/2011	3
Classement	1	LCEN	3
référencement	3	loi informatique liberté	2
taxe Google	3	directive 2001/29/CE	3

TAB. 1 – *Extrait du lexique constituant la classe 'hébergeur' dans notre ontologie*

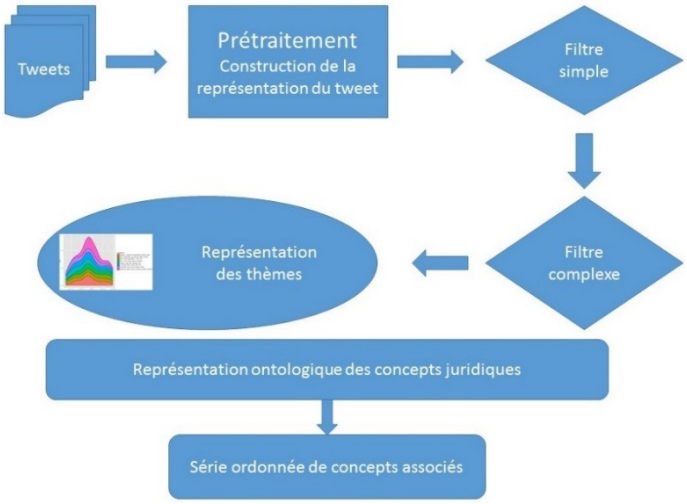


FIG. 1 – *Schéma de traitement par le système.*

4 Le système de traitement

Comme le montre la figure 1, notre système de traitement suit différentes étapes : un prétraitement, un passage par deux filtres avant la construction d'une représentation des thèmes qui sera comparée aux classes de concepts juridiques

4.1 Prétraitement

L'objectif est de rassembler les différentes informations portées par un Tweet organisées dans une structure qui sépare le texte et les méta-informations Les champs de stockages sont :

— Nature (Original, Retweet, Reply) : Nature du tweet. S'il est original, s'il est retweeté ou encore si c'est une réponse à un autre tweet.
— La date. La source.
— Liens : Liste des URLS, hastags et @users contenu dans le texte.
— Smiley : La liste des smileys contenus dans le tweet.
— Le contenu significatif : Le contenu significatif correspond au contenu textuel, après éliminations des caractères spéciaux (#, »,@), les accents, les symboles (comme le RT pour un « retweet »), les URL. Certains caractères sont conservés comme les « / » nécessaire pour identifier certaines mentions comme des règlements ou des directives. Il est ensuite confronté à une stop-liste pour retirer des termes considérés comme peu significatif. Finalement les termes sont 'lemmatisés' via un shallow parser, TreeTagger (Schmid (1994)) qui va simplifier le message en ôtant notamment la majorité des marqueurs grammaticaux. Le résultat est une forme canonique du message d'origine.

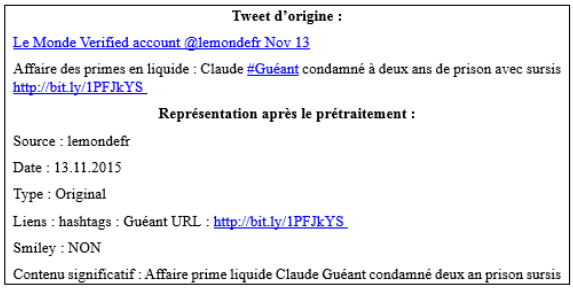


FIG. 2 – *Trace d'un tweet aprés son pretraitement*

4.2 Filtres

4.2.1 Filtre Simple

Le premier filtre est extrêmement simple, il élimine tous les tweets qui ne correspondent pas à des critères prédéfinis. Actuellement il y a deux critères d'élimination d'un tweet :

— Si le tweet ne contient aucun terme présent dans l'ensemble des lexiques rattachés aux classes de concept juridiques.
— Si le tweet contient des smileys.

Le second critère est en réalité dérivé du second filtre (cf. 4.2.2). Le second filtre est en effet basé sur des algorithmes de classification supervisée en machine Learning. Le message est alors dérivé en 'features' (éléments caractéristiques du message). L'une des features 'expr-emo' utilisait la présence de smiley. Or à l'intérieur de notre premier jeu de départ, aucun tweet pertinent ne contient de smiley. Ceux-ci étant en général tweetés dans un cadre 'professionnel' et francophone, le ton adopté se contient dans un registre soutenu.

Lorsqu'un descripteur prend le caractère d'une stop-liste, comme ici les smileys, son contenu (ou la partie pertinente) est transférée dans le filtre simple. Inversement si par évolution des données, elle cesse d'avoir une corrélation de 100

Ce filtre simple est responsable de la majorité des tweets rejetés par le système (rejet minimal sur nos test préliminaires : 95

4.2.2 Filtre Complexe

Le filtre complexe est un filtre entrainé sur corpus. Deux algorithmes présentant des résultats similaires (SVM et régression logistique) ont été implémentés en parallèle. Tous deux présentent des résultats accrus parmi les différents algorithmes testés (inductif, un système d'apprentissage de règles, SVM et logit) et un choix définitif pourra être fait suite à l'évaluation du système. Il s'inspire de Song (2012) où il a été utilisé dans la détection de signaux faibles (information à propos de la propagation de la grippe). L'algorithme d'apprentissage supervisé utilisé pour filtrer les tweets liés au domaine juridique repose sur un ensemble de features qui sont généralement binaires ou trivalentes. Un premier ensemble correspond à la

description de la forme du tweet : ainsi **hasurl, isretweet, isreply** ou **length** qui contient le nombre de caractère.

On a ensuite des features associées au lexique de notre ontologie.

Wordcount contient le nombre de termes référencés contenus dans le tweet. **Has_Signal** indique si le tweet contient un terme issu de l'ontologie. **Cont_2Sig_SameCategory** nous indique si le tweet contient au moins deux termes appartenant à la même classe. **Cont_2Sig-_DifCategory** indique si des termes différents existent qui appartiennent à au moins deux classes.

Correct_syntax_90 est une feature qui est encore en cours d'élaboration. Elle utilise un moteur de correction orthographique (actuellement celui en développement se base sur hunspell-fr) pour déterminer si le texte du tweet est syntaxiquement correct. En effet, les tweets 'professionnel' tendent à avoir une syntaxe impeccable. **Expr_Emo** indique si le tweet exprime une émotion, il se base sur un lexique contenant des verbes d'expression d'opinion comme aimer ou détester , des expressions colloquiale ou des termes plus ou moins soutenu. Cette feature contenait aussi des marqueurs tels que les smiley, mais ceux-ci sont actuellement trop porteur de sens pour faire partie d'une feature et ils ont été déplacé vers la stop liste du filtre simple. **Expr_slang** est très proches du **Expr_Emo**, elle filtre les termes populaires du langage propre à internet, tel que des acronymes ('lol' , 'mdr'). **Post_regular** indique si le posteur poste régulièrement ou non. La régularité du posteur s'est révélée significatives dans des analyses variée comme la discrimination entre un bot et un humain (Chu et al. (2012)) ou l'identification d'évènement Kunneman et v.d.Bosch (2014).

Finalement un dernier ensemble de feature qui proviennent de nos usages professionnels en tant que juriste tels que Contain_KnownSource, ou la source peut être le compte d'un grand cabinet d'avocat par exemple, cette feature permet d'établir facilement la corrélation entre une source d'information à caractère juridique et le contenu de ces tweets. Pour autant, un cabinet d'avocat par exemple peut très bien décider de faire une annonce à caractère évènementiel ('20e anniversaire du cabinet') ou publicitaire laquelle n'aura pas de pertinence. Dans le même ordre d'idée, **Contain_KnownTag, Comesfrom_Knownsource et Comesfrom_Knowntag** relient certains tweets avec leur contenu

Entrainement du filtre Notre jeu d'entrainement est pour le moment artificiel car nous n'avons pas encore récupéré un nombre de tweet francophone suffisant pour avoir des résultats solides. Nous avons donc extraitdes tweets juridiques que nous avons 'noyés' au milieu de tweets non pertinent. Environs 25 tweets pertinents pour 1000 non pertinents (soit 2,5L'ensemble permet de configurer un environnement plus ou moins homogène afin de faire quelques tests pour guider le processus de conception du système. Evidemment cela ne permet pas 'évaluer les performances du système face à une distribution non homogène d'évènement mais ce cas de figure requiert de toutes façons un système plus robuste donc entrainé sur un jeu plus conséquent. Sur ce jeu de test le 'Recall' (taux de tweet pertinent effectivement conservé parmi les tweets pertinents à détecter) atteint une valeur stable d'environ 80particulière d'un filtre pour signal faible, nous évaluons la précision comme étant le taux de tweet pertinents parmi les tweets qui 'traversent' le filtre. Les valeurs de précisions sont encore très variables en fonction de l'accroissement du jeu de données. Nous avons récemment entrepris de faire un accroissement incrémental du jeu de données, tant d'entrainements que celui de test. Pour contrôler cette augmentation en s'assurant d'un accroissement des performances du système,

on utilise le Framework ALIS Girard et al. (2015) pour permettre au système de nous indiquer les tweets à catégoriser en priorité afin d'optimiser le gain de performance du système dans le temps.

4.3 Représentation des thèmes et analyse de la classe appropriée

Les différents thèmes doivent être représentés afin de pouvoir les comparer avec le contenu des classes.

Les thèmes (topic) sont une fonction des éléments de vocabulaire contenu dans le texte du tweet mais aussi de son fil. Pour effectuer un classement on utilise des techniques non supervisé de reconnaissance de la salience des termes employés tel que présenté par Zhao (2013). Lorsque l'information est disponible, le système va prendre en compte n tweets voisins dans le fil, n'étant un paramètre que l'on fait évoluer en fonction des résultats pour calibrer cette représentation. Une fois cette représentation effectuée, on effectue une recherche de cluster sommaire, il n'y a rarement plus de quelques termes commun a un tweet, le clustering est alors presque effectué sans effort.

Pour permettre de relier les thèmes associés au tweets à nos classes on se base sur un modèle de traitement latent, LDA (Latent Dirichlet Allocation). Suivant (Pépin (2015)), LDA étant un modèle probabiliste à variable latente, il suppose l'existence d'une variable cachée qui conditionne l'existence de termes dans les documents : les thèmes. Les paramètres du modèle sont le nombre de thèmes et les paramètres des distributions a priori, appelés hyper-paramètres. La valeur de l'un de ces hyper-paramètres régule leur répartition sur les thèmes : des composantes faibles auront tendance à forcer les documents à être décrits par peu de thèmes, alors que des valeurs élevées auront tendance à donner une description plus répartie des documents. De manière analogue, un second hyper-paramètre échoit à la description des thèmes et régule leur répartition sur les termes du vocabulaire : des valeurs faibles de ce paramètre auront tendance à produire des thèmes spécifiques, décrits par peu de termes, alors que des valeurs plus élevées auront tendance à créer des thèmes plus homogènes, décrits par un nombre plus importants de termes. Par ailleurs, même si le modèle est décrit en utilisant le vocabulaire de la fouille de textes , (en raison de son origine), son application n'est pas restreinte à ce domaine.

LDA s'applique à n'importe quelle variable catégorielle, et a été adapté aussi bien à l'analyse d'images (Liénou et al. (2010) ; Fei-Fei et Perona (2005)), qu'à la bio-informatique (Flaherty et al. (2005)) ou aux systèmes de recommandation (Blei et al. (2003)).

Dans le domaine du de la fouille de données textuelles, différentes variantes de la LDA sont régulièrement utilisés. L'une de ces variantes proposées par Samdani et al. (2014) offre un modèle permettant de confronter l'apprentissage à des connaissances a priori pour créer une série de contrainte permettant la résolution de coréférence. L'approche L3M (tuned gamma) permet d'implémenter une recherche de cluster fondé sur les topics. Dans une ligne similaire, Pépin et al. (2015), proposent une méthode d'analyse de topic fondé sur la LDA. L'usage de la normalisation par la LDA et sa capacité à représenter des clusters comme des éléments continus (clustering flou), permet de faire apparaitre des concepts comme thèmes qui serait passé inaperçus sur un jeu de données préfiltrés comme le nôtre, lequel serait autrement apparu homogène, ce qui permet de faire le lien avec notre ontologie des concepts du droit. L'intérêt de cette d'approche étant la possibilité d'utiliser les distributions de probabilités pour calculer une similarité

L'action de la LDA permet effectivement de définir une forme de métrique entre une classe et une instance de tweet. Toutefois cette métrique ne saurait être précise tant qu'une méthode de répartition des coefficients associés à chaque concept n'a pas été déterminée. Les coefficients ad hoc que l'on a attribués ne reflètent pas la réalité en termes de pondération d'importance des concepts.

L'usage d'un model entrainé sur un gros jeu de données issu de Twitter entrainent des problèmes de traitements et de stockages particuliers propres aux problématiques du BIG DATA (Tjong Kim Sang et van den Bosch (2013)).

Pour le traitement de données dans le temps par la LDA, Lau et al. (2012) ont remplacé la factorisation matricielle (qui aurait impliqué l'ensemble des données) par un système de traitement itératif ou les paramètres d'un jeu recevant de nouvelles données sont héritées du model issu de l'ancien jeu de données. Ce procédé permet une approche dynamique du traitement par la LDA (Lau, Collier et Baldwin, 2012) approprié pour notre système supposé effectuer une veille, ce qui implique une constante évolution du jeu de donnée.

4.4 Traitement des URLS

Lorsqu'un tweet contient une URL , on effectue un traitement similaire a celui du tweet sur le texte de la page. On exclue pour le moment les liens multimedia (liens vers une vidéo par exemple). On traitera quand meme le texte présent dans le corps de la page introduisant une vidéo pour certains liens contenant une vidéo *"embedded"*. Dans un soucis de simplicité , on soumet le texte obtenu au même traitement par les filtres afin de determiner la pertinence. On obtient ainsi en traitant la page web, un accroissement du Recall d'environ 5%.

Cette approche semble suffisante dans une optique d'evaluation de l'intêret même si elle souleve quelques inconveniences mineures et facile à resoudre dans un futur proche tel que :

— Des calculs inutiles sur des procédures pertinentes pour un tweet et non pertinente pour une texte. Il conviendra donc de tracer un process plus traditionnel de filtre textuel.

— Quelques faux positifs dûts au fait que les filtres etant prévus pour des tweets , la simple occurence fait foi en terme descripteur , la ou la densite prend plus de sens pour la description d'un texte.

— Les descripteurs du filtre complexe lié aux style ne sont plus aussi informatifs. Il conviendra donc de redefinir les descripteurs. Par contre un filtre complexe entrainé sur un fil de tweet parvient à fonctionner correctement sur une page-web linkée , ce qui n'est pas un résultat intuitif a priori même si la repartition des vecteurs d'etats d'ALYS le demontrait.

Donc l'addition du traitement d'URL offre de bonnes perspectives au vu des premiers essais ,donc nous prevoyons la mise au point d'un traitement approprié.

4.5 Conclusions et travaux futurs

La description du système dans cet article présente la 'roadmap' que nous suivons pour la construction d'un premier prototype du système : mise au point des filtres destinés à isoler des signaux faibles, construction d'une ontologie efficace du droit, définition d'une métrique permettant d'établir les similarités entre signaux. Mais pour pouvoir réellement atteindre nos objectifs, et avoir un outil véritablement efficace, un travail de profondeur reste à faire, notamment en philosophie appliquée afin de déterminer ce que signifie construire une ontologie de

concepts dans un domaine comme le droit. Un important travail juridique est aussi nécessaire pour construire les classes que constitue cette ontologie. Une sélection de textes doit être menée puis ceux-ci doivent être analysés, probablement par un expert autant que par des algorithmes. La partie 'data science' est constituée de choix qui ne sont que prospectif quand il ne s'agit pas simplement de gouts. Ils ne sauraient être fixés dans le marbre. Ainsi le 'filtre complexe' utilise un algorithme supervisé heureusement peu gourmand en données d'apprentissage. En l'espèce, les différents descripteurs définies pourraient tout aussi bien servir à un apprentissage de type non-supervisé. Même la combinaison d'algorithme non supervisé afin de faire la jonction entre les classes, bien que fort prometteuse, est encore en phase de test. Il conviendra aussi de définir une méthode de sélection et d'apprentissage automatisé des constituants des classes à partir des sources du droit. Si on se projette assez loin, il va aussi se poser des questions d'ingénierie informatique. Les connaissances a priori étant fonctions de l'actualité, leur révision devra forcement prendre la forme de mise à jour régulière comme c'est déjà le cas pour de nombreux filtres appliqués à internet, tels les filtres anti-spam ou les anti-virus. L'objectif de cet article est de servir d'inspiration et d'échanges d'idées pour des chercheurs s'intéressant à des sujets similaires, d'ouvrir la possibilité de coopérations sur un des nombreux * aspect disciplinaire du projet ou plus simplement de permettre une discussion et des remarques au sein de l'atelier de fouille de données complexes dont la problématique est d'une importance g randissante, notamment pour des jeux de données francophones

Références

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 14*, 993-1022, 2003. 19, 32, 35, 70, 72, 74, 76, 92, 102.

Chu, Z., S. Gianvecchio, H. Wang, et S. Jajodia (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing 9*, 811–824.

CJUE (2015). Libre prestation des services, l'offre de courtes vidéos sur le site internet d'un journal peut relever de la réglementation des services de médias audiovisuels. *arrêt du 21 octobre 2015, New Media Online, C-347/14. Communiqué de presse no 127/2015 du 21/10/2015*.

Delmas-Marty, M. (1986). *Le flou du droit*. Edition presse universitaire de France.

Fei-Fei, L. et P. Perona (2005). A bayesian hierarchical model for learning natural scene categories. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference 2*, 524–531.

Flaherty, P., G. Giaever, J. Kumm, M. I. Jordan, , et A. P. Arkin (2005). A latent variable model for chemogenomic profiling. *Bioinformatics 21*, 3286–3293.

Girard, Y., P. Saurel, et F. Rousseaux (2015). Facilitating large data analysis in virtual humanities and e-science: The alis paradigm. *CIE45 proceedings 45*.

Harryson, M., E. Métayer, et H. Sarrazin (2014). The strength of "weak signals". *McKinsey Quarterly*.

JO (2015). Ordonnance no 2015-948 du 31 juillet 2015 art.15. *Journal officiel de Août 1978 7*, 28–32.

Kunneman, F. et A. v.d.Bosch (2014). Automatically identifying periodic social events from twitter. *proc. Recent Advances in Natural Language Processing.*.

Lau, J., N. Collier, et T. Baldwin (2012). On-line trend analysis with topic models: #twitter trends detection topic model online. *Proceedings of COLING 12*, 1519–1534.

Liénou, M., H. Maître, et M. Datcu (2010). Semantic annotation of satellite images using latent dirichlet allocation. *Geoscience and Remote Sensing Letters, IEEE 7*, 28–32.

Loi, L. (2010). Loi no 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique.

Ordonnance (2015). Ordonnance no 2015-948 du 31 juillet 2015 art.15. *Journal officiel du 2 aout 2015 7*, 28–32.

Pépin, L. (2015). *Fouille exploratoire de messages publiés sur Twitter pour l'aide á la décision*. Thèse de doctorat, université de Nantes.

Pépin, L. et al. (2015). Visual analytics for exploring the topic evolution of company targeted tweets. *CIE45 proceedings*.

Pierre, S. (2014). *Ingénierie des connaissances pour traiter de l'hétérogénéité des données issues de systèmes sociaux, IA et altérite radicale*. Thèse d'habilitation à diriger des recherches, Université de Reims Champagne Ardenne, Ecole Doctorale Sciences Technologies Santé.

Power, D., J. Daniel, et P.-W. Gloria (2011). Impact of social media and web 2.0 on decision-making. *Journal of decision systems 20(3)*, 249–261.

Samdani, R., K.-W. Chang, et D. Roth (2014). A discriminative latent variable model for online clustering. *Proceedings of the 31st International Conference on Machine Learning 32*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the international conference on new methods in language processing 12*, 44–49.

Sirinelli, P., A. Bensamoun, et C. Pourreau (2014). Rapport de la mission sur la révision de la directive 2001/29/ce sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information. *rapport du conseil supérieur de la propriété littéraire et artistique*.

Song, B. (2012). *Weak Signal Detection on Twitter Datasets: A non-accumulated approach for non-famous events*. Master thesis, TU Deft department of Software Technology.

Tjong Kim Sang, E. et A. van den Bosch (2013). Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal 3*, 121–134.

Zhao, Y. (2013). Topic modelling and social data analysis in r and data mining. *Elsevier Edition*, 121–134.

Summary

This article presents methodological directions in designing a system capable to learn the concept of law and use this knowledge to analyse an information flux seachin for relevant information.

Identification de Relations Sémantiques Supportée par des Algorithmes d'Apprentissage Supervisé Utilisant des Mesures de Proximité Sémantique

Jean Petit[*,**] Jean-Charles Risch[*,**,***]

[*]Université de Reims Champagne Ardenne, CRéSTIC EA 3804, Moulin de la Housse, 51100 Reims
jean.petit, jean-charles.risch@etudiant.univ-reims,
http://crestic.univ-reims.fr/
[**]Capgemini Technology Services, 7 rue Frédéric Clavel, 92287 Suresnes
jean.petit, jean-charles.a.risch@capgemini.com
https://www.fr.capgemini.com/
[***]Université Technologique de Troyes, 12 rue Marie Curie, Tech-CICO, 10010 Troyes
http://www.utt.fr/

Résumé. Les relations sémantiques sont au cœur d'un nombre grandissant de systèmes à base de connaissances. Le besoin de découvrir de telles relations a mené à des avancées concluantes dans leur acquisition automatique. Cependant, leur validation automatique reste un défi d'actualité. Dans cet article, nous présentons une méthode permettant d'identifier automatiquement la classe d'une relation sémantique à partir du web. L'hypothèse initiale est qu'il est possible de déterminer la proximité entre des relations sémantiques et des classes sémantiques en ajoutant dans l'expression des mesures de similarité, des motifs syntaxiques indicatifs de classes sémantiques. En utilisant ces motifs comme variables descriptives d'entrées pour des algorithmes d'apprentissage, nous sommes ainsi capables d'identifier des relations sémantiques. Nos expérimentations ont conduit à labelliser correctement 72 % de relations sémantiques réparties dans trois classes sémantiques : hyperonyme, méronyme et causal, dans un environnement sémantique riche avec une classe mixte contenant des antonymes, synonymes et co-hyponymes.

1 Introduction

Les relations sémantiques, telles que les relations d'hyperonymie, de méronymie ou encore de causalité, sont au cœur de nombreuses bases de connaissances qui supportent chaque jour de plus en plus d'applications. Leur découverte et leur évaluation représentent ainsi un défi majeur. Si l'acquisition des relations sémantiques peut être faite de manière automatique, leur validation reste une tâche manuelle, laborieuse et sujette aux erreurs.

Les mesures de similarité sont des indicateurs adaptés pour estimer la proximité de deux entités ou paires de mots en se basant sur leur cooccurrence. Ces mesures s'ancrent sur l'hypothèse distributionnelle harissienne (Harris, 1954) qui suppose que les mots « proches » ap-

paraissent dans des contextes similaires. Les moteurs de recherche tels que Google, Yahoo, Bing ou encore Yandex offrent un environnement adéquat pour le calcul de cette mesure en répondant aisément aux besoins quantitatif et qualitatif. En fournissant un accès simplifié au web au travers de leur API réciproque, il est ainsi possible d'exploiter une source très large et diverse de données textuelles de différentes temporalités. Les informations retournées par les moteurs de recherche sont la source de trois types de méthode pour calculer une mesure de similarité. Ces méthodes sont basées sur :

— le nombre de résultats retournés (Gledson et Keane, 2008) ;
— les extraits de texte (Sahami et Heilman, 2006) ;
— les deux à la fois (Bollegala et al., 2007).

Néanmoins, si les mesures de similarité suggèrent l'existence d'une relation sémantique entre deux entités, elles n'apportent pas la dimension sémantique nécessaire pour identifier une potentielle classe sémantique.

Hearst (Hearst, 1992) développe une méthode qui permet d'exploiter des motifs syntaxiques afin d'extraire des relations d'hyperonymie. Par exemple, le motif « Y est X » avec X et Y des locutions nominales, permet de découvrir dans la phrase « Un chien est un animal domestique. » qu' « animal domestique » est l'hyperonyme de « chien ». Elle prouve ainsi que certains motifs syntaxiques sont indicatifs de cette relation sémantique particulière. Par la suite, d'autres chercheurs (Girju et Moldovan, 2002, 2003, 2006; Caraballo, 1999; Cederberg et Widdows, 2003; Pantel et Ravichandran, 2004) appliquent avec succès sa méthode à d'autres relations sémantiques et ainsi la généralisent. Cependant, la fréquence et la qualité de ces motifs syntaxiques plutôt rares varient, il est donc difficile de juger de la validité des relations sémantiques extraites.

Oliviera (De Oliveira, 2009) inclut la dimension sémantique dans l'expression de la mesure de similarité en y associant des motifs syntaxiques. Une mesure de proximité sémantique est ainsi calculée en se basant sur la cooccurrence des entités qui composent une relation sémantique ainsi qu'un motif syntaxique représentatif d'au moins une classe sémantique. Pour avoir des mesures statistiques significatives, les motifs syntaxiques étant rares rappelons-le, Costa et al. (Costa et al., 2011) ont utilisé une approche basée sur le web, leur donnant ainsi accès à une large quantité de données. Leur méthode compile les différentes mesures de proximité sémantique associées à différents motifs syntaxiques indicatifs d'une même classe sémantique afin d'avoir une unique mesure, ce qui leur permet avec un seuil de séparer les relations sémantiques potentiellement bonnes des mauvaises. S'ils affichent un très haut taux de réussite dans un environnement relativement simple : distinction entre relations correctes (sémantiques valides) et incorrectes (non sémantiques). Leur méthode révèle cependant des faiblesses pour distinguer entre elles des relations sémantiques correctes associées à différentes classes sémantiques.

Notre objectif principal est d'identifier la classe de relations sémantiques. Dans ce papier, nous supposons que les mesures de similarité générées à partir de motifs syntaxiques permettent de décrire la proximité sémantique entre une relation et une classe. Par conséquence, la classe sémantique d'une relation se traduit par des mesures caractéristiques associées à certains motifs syntaxiques. Ainsi, nous supposons qu'il est possible en utilisant des algorithmes de classification de reconnaître la classe sémantique de relations prises au hasard.

Mis à part la réalisation de l'objectif principal de notre étude, nos contributions sont les suivantes :

— Une étude sur diverses contraintes contextuelles associées à l'expression des mesures de proximité sémantique.
— Le développement d'un score qui se base sur une vérification de la cohérence entre la requête soumise au moteur de recherche et les résultats retournés afin d'améliorer la qualité des mesures de proximité sémantique.
— Une analyse comparative de la performance de trois algorithmes d'apprentissage que sont l'arbre de décision, la forêt d'arbres décisionnels et le réseau de neurones.

L'expérience menée dans ce papier démontre que notre méthodologie évalue correctement 72 % des relations sémantiques testées. C'est un score encourageant pour l'évaluation de relations d'hyperonymie, de méronymie et de causalité dans un environnement sémantique bruité par des relations de synonymie, d'antonymie et de co-hyponymie. Sans ce bruit, le pourcentage de relations correctement identifiées atteint 82 % et 90 % en portant l'analyse uniquement sur la distinction entre hyperonyme et méronyme. Ces résultats nous rapprochent encore un peu plus d'une possible évaluation automatique de relations sémantiques.

Il est important de spécifier que nous travaillons avec la langue anglaise. Cependant, la méthodologie développée dans ce papier est applicable à d'autres langues.

Dans un premier temps nous expliquons le choix du moteur de recherche sur lequel se fonde notre méthode. Puis, nous présentons le calcul de mesures de proximité sémantique basé sur le web et introduisons des pistes d'amélioration. Ensuite, nous développons la partie concernant l'apprentissage des classes sémantiques. Nous exposons notre méthodologie ainsi que l'expérience qui s'en suit. Enfin, nous discutons les résultats obtenus et concluons en ouvrant sur des perspectives d'évolution.

2 Importance du Choix du Moteur de Recherche

Les méthodes avec des approches basées sur le web sont dépendantes des moteurs de recherche qu'elles utilisent, leurs résultats aussi. C'est pourquoi il est nécessaire de choisir son moteur avec soin et en connaissance de cause. Les plus cités dans les papiers scientifiques qui nous concernent sont dans l'ordre : Google, Yahoo, puis Bing. Cependant, est-ce pour autant ceux qui répondent le mieux à nos besoins ? L'objectif ici n'est pas de faire une étude comparative des différents moteurs de recherche qui sont sur le marché, mais d'expliquer le choix du moteur de recherche Yandex pour notre étude.

Il y a de nombreux paramètres à prendre en considération lors de l'évaluation d'un moteur de recherche. Tout d'abord, il est nécessaire qu'une API permette la communication entre une application et ce dernier. Ensuite, il faut qu'il offre l'accès à des informations de base, à minima le nombre de résultats retournés à la suite d'une requête ainsi que les extraits de texte correspondants. Le volume de données stockées est aussi important puisque le moteur est exploité comme une passerelle pour accéder à une dimension Big Data. D'autres facteurs limitant comme la taille maximum de caractères dans la requête ainsi que les opérateurs logiques conditionnent les interactions possibles avec le moteur et sont autant de restrictions associées à la méthode. Finalement, le coût des requêtes est souvent décisif dans la mise en place des solutions tournées vers le web.

En se basant sur ces différents critères, nous nous sommes tournés vers Yandex qui est un moteur de recherche russe. Il possède un nombre de pages indexées en anglais (qui est la langue avec laquelle nous travaillons) similaire à Google. Comme ce dernier, il permet d'effectuer des

recherches prenant nativement en compte la forme lemmatisée des mots, ce qui nous permet de faire abstraction en partie de la gestion des déclinaisons, gestion souvent impossible du fait de la limite associée au nombre de caractères présents dans la requête. Il offre une large gamme d'opérateurs logiques [1] qui le démarquent grandement des divers leaders du marché. Il est par exemple possible de gérer avec exactitude la distance qui sépare les différents mots qui composent une requête : « $crime\ /-1\ and\ /(+2\ +5)\ car$ » signifie que nous cherchons un mot dont le lemme est « crime » accolé à gauche à un mot dont le lemme est « and » et avec un mot dont le lemme est « car » entre 2 et 5 mots après. Dernièrement, l'aspect économique a été décisif puisque Yandex offre un nombre de requêtes automatiques gratuites qui permet de réaliser des expériences relativement librement sur le web. En effet à la date du 10/12/2015, si Google en offre 100 par jour, Bing 5000 par mois (environ 165 par jour), Yandex en offre 10000 par jour.

3 Les Mesures de Proximité Sémantique Basées sur le Web

Une relation sémantique se caractérise par un triplet (e_1, r, e_2) où $e_1$et e_2 sont des entités (unités lexicales) et r la relation. Calculer la proximité des entités e_1 et e_2 en utilisant un moteur de recherche est possible en effectuant plusieurs requêtes Q permettant d'accéder au nombre de résultats P associés à $Q(e_1), Q(e_2)$ et $Q(e_1 \cap e_2)$. Les formules les plus populaires pour calculer ces mesures sont Web Jaccard, Web Overlap, Web Dice, Web PMI (Pointwise Mutual Information) ou encore NWD (Normalized Web Distance). On peut trouver une description plus complète sur le papier d'Islam et al. (Islam et al., 2012).

Costa et al. (Costa et al., 2011) ont changé l'expression initiale dédiée au calcul des mesures de similarité pour inclure la dimension sémantique au travers de motifs syntaxiques πri. Ainsi, les requêtes requises pour calculer la mesure de proximité sémantique entre deux locutions nominales en fonction d'un motif syntaxique sont :

— $P(e_1) = Q(e_1\ /+1\ \pi ri)$;
— $P(e_2) = Q(\pi ri\ /+1\ e_2)$;
— $P(e_1 \cap e_2) = Q(e_1\ /+1\ \pi ri\ /+1\ e_2)$.

Par exemple, la mesure de proximité sémantique pour la relation « crime, robbery » associée au motif syntaxique « such as » indicatif d'une relation d'hyperonymie est obtenu en récupérant les valeurs :

— $P(e_1) = Q(crime\ /+1\ such\ /+1\ as)$;
— $P(e_2) = Q(such\ /+1\ as\ /+1\ robbery)$;
— $P(e_1 \cap e_2) = Q(crime\ /+1\ such\ /+1\ as\ /+1\ robbery)$.

Les mesures de proximité sémantique permettent ainsi d'indiquer l'affinité d'une relation avec un motif syntaxique et donc par extension de définir une probabilité d'appartenir à une classe sémantique particulière.

3.1 Maîtriser le Contexte

« Les mesures de cooccurrence basées sur le nombre de résultats retournés [. . .] ne prennent pas en compte le contexte local dans lesquelles [deux entités] cooccurrent » (Haribabu et al.,

1. Vous pouvez consulter la description des opérateurs logiques sur cette page web : https://yandex.com/support/search/how-to-search/search-operators.xml.

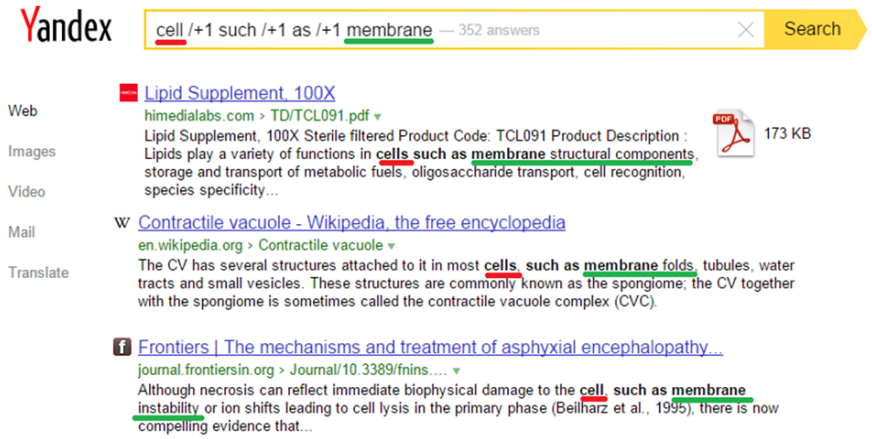


FIG. 1 – *Exemple de requête pour les entités « cell » et « membrane » avec le motif syntaxique « such as » effectuée sur le moteur de recherche Yandex.*

2014). Des mots ayant plusieurs sens porteront ainsi avec eux cette ambiguïté qui se reflétera dans le nombre de résultats retournés (Bollegala et al., 2010). C'est-à-dire que pour une recherche avec le mot « orange », le total retourné va correspondre à la fois au fruit, à la couleur, mais aussi à l'opérateur de téléphonie.

Dans l'expression originale des mesures de proximité sémantique, aucune contrainte n'est imposée au contexte bien que rien ne l'empêche. Les opérateurs logiques de Yandex autorisant la gestion de la taille du contexte, nous avons décidé de saisir cette opportunité pour étudier deux niveaux de contexte : dans le même document et dans la même phrase. Les formules correspondantes sont :

— $P_s(e_1) = Q(e_1 \ / + 1 \ \pi ri \ \& \ e_2) | P_w(e_1) = Q(e_1 \ / + 1 \ \pi ri \ \&\& \ e_2)$
— $P_s(e_2) = Q(\pi ri \ / + 1 \ e_2 \ \& \ e_1) | P_w(e_2) = Q(\pi ri \ / + 1 \ e_2 \ \&\& \ e_1)$

3.2 Vérifier la Cohérence des Résultats

Les mesures sémantiques sont basées sur des « requêtes intelligentes » qui ont pour but de faire émerger leur relation avec une classe sémantique par association à un motif syntaxique. Cela repose sur l'idée que les entités introduites dans la requête sont bien celles qui vont être à l'origine du nombre de résultats retourné par le moteur de recherche. Cependant les moteurs de recherche effectuent une simple correspondance mot à mot. Il ne traite pas des locutions nominales.

Par exemple, en observant la figure 1, le nombre de résultats associé est de 352 alors que de nombreux résultats correspondent à des relations sémantiques dont la locution nominale n'est pas « membrane ». Pour contourner ce problème, des chercheurs utilisent des motifs syntaxiques à double ancrage (Kozareva et al., 2008) comme « e_1 such as e_2 and ». Cependant, cette méthode n'est pas appropriée pour chaque motif syntaxique et elle diminue significativement le nombre total de résultats.

Pour gérer ce problème, nous avons développé un score de « conformité » qui permet d'estimer la cohérence entre les locutions nominales présentes dans la requête et celles retrouvées dans les extraits de texte. Lors du calcul de proximité sémantique, nous récupérons le maximum de 100 extraits de texte retournés avec la requête $Q(e_1 \cap e_2)$. Pour chaque extrait, la classe grammaticale des mots est trouvée à l'aide de TreeTagger [2] ce qui permet de définir des locutions nominales simples. La construction d'expressions régulières basées sur le motif syntaxique présent dans la requête web permet l'extraction des locutions nominales adjacentes. Pour chaque extrait de texte, on compare la relation sémantique recherchée (e_1, r, e_2) avec celles extraites (e_x, r, e_y). Si les entités e_1 et e_2 sont identiques ou correspondent à la tête respectivement de e_x ou e_y, alors on considère que l'extrait de texte est en accord avec la recherche, dans le cas contraire on retient l'incohérence. Finalement, le score de conformité est calculé en se basant sur le nombre d'extraits valides par rapport au nombre total d'extraits. Ce score sera ensuite utilisé afin de pondérer le nombre de résultats retournés et ainsi s'approcher un peu plus de la réalité.

4 Apprentissage des Classes Sémantiques

Nous avons vu dans la partie précédente la façon dont nous construisons les caractéristiques représentatives d'une relation sémantique. Nous rappelons que notre objectif est l'identification automatique des classes sémantiques de ces relations. Afin d'y parvenir, nous avons eu besoin d'utiliser des techniques d'apprentissage, notamment automatiques.

L'apprentissage automatique se divise en deux grands sous domaines : l'apprentissage supervisé et l'apprentissage non supervisé. Dans le premier cas, il est question d'apprendre un modèle statistique à partir de données déjà labellisées. Dans le second, il est question de construire les classes (inconnues par la machine) à partir des données.

Dans notre cas, nous connaissons déjà les classes que l'on cherche à identifier. Ce contexte est adapté aux conditions de l'apprentissage supervisé. De plus, les relations sémantiques sont décrites par des valeurs numériques continues. Plusieurs genres d'algorithmes correspondent à nos besoins : les réseaux de neurones, les arbres de décisions (Breiman et al., 1984) ou encore les forêts d'arbres décisionnels (Breiman, 2001) pour les plus connus. Ce sont donc ces algorithmes que nous utiliserons lors de notre expérience.

5 Expérience

L'objectif principal de notre expérience est d'évaluer l'efficacité des diverses mesures sémantiques permettant d'identifier la classe sémantique d'une relation sémantique. Nous avons décidé de comparer trois algorithmes d'apprentissage supervisé ainsi que d'évaluer l'impact des différentes manières de produire ces mesures en fonction de la présence ou l'absence d'un score de conformité ainsi que le type de contrainte associée au contexte (aucun, phrase, page web). Notre base de référence est une mesure sans pondération ainsi que sans contrainte sur le contexte. A l'aide d'une validation croisée de type « k-fold », nous avons calculé le taux d'erreur des modèles statistiques. Notons que, concernant la validation croisée, nous avons

2. Pour plus d'information sur ce POS tagger, vous pouvez consulter cette addresse : http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

sélectionné $k = 10$ comme il est d'usage dans la communauté. Nous les avons finalement comparés entre eux pour découvrir la meilleure solution.

5.1 Présentation des Données d'Entrée

Dans ce papier, nous nous sommes focalisés sur l'identification de trois classes sémantiques principales : hyperonymie, méronymie et causalité, ainsi qu'une quatrième classe composée de relations sémantiques qui n'appartiennent pas à ces trois dernières. Notre ensemble de données est composé de 148 hyperonymes, 200 méronymes, 179 relations causales et un « pot-pourri » de 182 relations sémantiques mixtes (synonymes, antonymes et co-hyponymes). Wordnet [3] a été utilisé pour extraire une partie des relations sémantiques, l'autre partie a été ajoutée manuellement. A l'origine, chaque classe était composée de 200 relations, la variation sur le nombre total final étant due à un filtre qui exclut les triplets (e_1, r, e_2) avec une cooccurrence inférieure à 10000. Nous avons estimé que de tels triplets ne sont statistiquement pas pertinents pour notre étude.

5.2 Sélection des Motifs Syntaxiques

Classes sémantiques	Motifs syntaxiques
Hyperonyme	e_1 *such as* e_2
	e_1 *like* e_2
	e_2 *and other* e_1
	e_2 *or other* e_1
Méronyme	e_1 *have* e_2
	e_2 *in* e_1
	e_2 *of* e_1
	e_2 *from* e_1
	e_2 *part of* e_1
Causalité	e_1 *cause* e_2
	e_1 *induce* e_2
	e_1 *generate* e_2
	e_1 *produce* e_2
	e_1 *lead to* e_2
	e_2 *due to* e_1

TAB. 1 – *Classes sémantiques et motifs syntaxiques associés.*

Les précédentes études ont permis la découverte d'un grand nombre de motifs syntaxiques indicatifs de l'hyperonymie (Hearst, 1992), de la méronymie (Berland et Charniak, 1999) et de la causalité (Girju et Moldovan, 2002). Le tableau 1 liste les motifs syntaxiques des classes sémantiques que l'on souhaite identifier dans notre expérience. Leur sélection s'est faite selon leur fréquence et leur ambiguïté qui devaient être respectivement forte et faible.

3. Disponible sur le site : https://wordnet.princeton.edu/

5.3 Configuration des Algorithmes d'Apprentissage

Nous avons testé trois algorithmes d'apprentissage supervisé : l'arbre de décision, la forêt d'arbres décisionnels et le réseau de neurones. Dans cette sous partie, nous présentons la configuration des paramètres les plus importants de ces algorithmes. Etant donné que nous nous plaçons dans un environnement statistique, nous allons nommer les relations sémantiques « individus » et les 15 motifs syntaxiques qui les décrivent « variables ».

5.3.1 Arbre de Décision

L'arbre de décision est un algorithme simple à configurer. Le principal paramètre à configurer est le critère permettant la sélection de la variable séparant au mieux les individus d'une classe. Il est usuel de choisir entre l'indice de Gini ou l'entropie. Le choix dépend généralement des données d'entrée, ainsi nous avons sélectionné le critère réduisant au mieux l'erreur de classification. Dans notre cas, nous avons sélectionné l'indice de Gini.

Le package R utilisé est rpart [4].

5.3.2 Forêt d'Arbres Décisionnels

L'algorithme de la forêt d'arbres décisionnels est un algorithme dérivé de l'arbre de décision. En effet, il est question de créer n arbres de décision calculés à partir de n sous-ensemble de données pris au hasard (individus et variables). Sa configuration se concentre autour du nombre n d'arbres à sélectionner et du critère de sélection des variables les plus discriminantes. Pour les mêmes raisons que nos choix sur l'arbre de décision, nous avons sélectionné l'indice de Gini. (Oshiro et al., 2012) explique qu'augmenter le nombre d'arbres ne signifie pas forcément améliorer les résultats. Il est conseillé de faire des tests de ce paramètre jusqu'à observer une certaine stabilité des résultats. Ainsi, nous avons testé 5 valeurs de ce paramètre (10, 30, 50, 70 et 100). Nous avons remarqué une amélioration des résultats jusqu'à une stabilité de ces derniers pour n : 70. Nous avons donc choisi cette valeur de n pour la configuration de la forêt d'arbres décisionnels.

Le package R utilisé est randomForest [5].

5.3.3 Réseau de Neurones

L'algorithme de réseau de neurones est une méthode d'apprentissage qui demande une configuration au préalable sur-mesure, essentielle à son bon fonctionnement (Lippmann, 1987). C'est pourquoi nous allons l'expliquer en détail.

La couche d'entrée du réseau de neurones ayant autant de nœuds que les individus ont de variables descriptives, notre réseau de neurones en comporte donc $n = 15$.

Nous cherchons à identifier 4 classes sémantiques : hyperonymie, méronymie, causalité et une dernière classe composée de relations sémantiques n'appartenant pas aux trois premières. La couche de sortie du réseau de neurones contenant autant de nœuds que de classes à identifier, notre réseau de neurones en compte donc quatre.

4. Voir https://cran.r-project.org/web/packages/rpart/rpart.pdf
5. Voir https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Enfin, ayant fait le choix d'un perceptron monocouche, nous avons une unique couche cachée. Le nombre de neurones de la couche cachée dépend de plusieurs facteurs dont le nombre de neurones de la couche d'entrée, le nombre de neurones de la couche de sortie ou encore le nombre d'individus à analyser. Les formules indicatives prescrites par (Tufféry, 2012) pour trouver le nombre de neurones de la couche cachée d'un perceptron monocouche sont :

$$pop = 5m(n+k)$$
$$n/2 < m < 2n$$

Avec n le nombre de neurones de la couche d'entrée, k le nombre de neurones de la couche de sortie, m le nombre de neurones de la couche cachée et pop le nombre d'individus à analyser.

Dans notre étude, nous avons $pop = 709$; $n = 15$ et $k = 4$. Le résultat de l'équation proposée est $m = 7.46$. Cependant l'inéquation indique que m doit être supérieur à $n/2$ soit 7.5. Nous n'avons donc pas de solution possible pour m dans notre étude. 8 étant la première valeur entière acceptable, c'est celle que nous avons sélectionnée pour nos expériences.

Notons que nous avons fait varier cette variable dans le but de l'ajuster en fonction des résultats obtenus. Cependant, ces derniers différaient peu d'une valeur de variable à l'autre. Ainsi, nous avons conservé ce nombre de 8 nœuds dans la couche cachée.

Enfin, il est également possible de modifier la fonction de transfert des nœuds d'un réseau de neurones. Dans notre étude, nous avons utilisé la fonction logistique car elle est la fonction la plus utilisée par les experts. D'autres fonctions de transfert ont été testées mais aucune amélioration n'a été observée sur les résultats.

Le package R utilisé est nnet [6]

6 Évaluation des Modèles Statistiques

Pour estimer la qualité de notre méthodologie, nous avons construit vingt-quatre modèles statistiques, un pour chaque analyse combinant le choix d'un algorithme d'apprentissage, le choix d'un contexte (page web, phrase ou aucun) et la présence ou non du score de conformité. Nous avons décidé de comparer deux formules en se basant sur les résultats obtenus par (Costa et al., 2011). Ces formules sont Web Dice et Web PMI. Ci-après un rappel de ces formules avec N le nombre total de pages indexées sur le moteur de recherche.

$$WebDice(e_1, e_2) = \frac{2*P(e_1 \cap e_2)}{P(e_1)+P(e_2)}$$
$$WebPMI(e_1, e_2) = log_2(\frac{P(e_1 \cap e_2)}{P(e_1)*P(e_2)} * N)$$

Les résultats de notre expérience sont synthétisés par le diagramme (figure 2). Nous avons observé que le réseau de neurones était l'algorithme d'apprentissage permettant d'obtenir le plus petit taux d'erreur, et ce malgré la régularité de bons résultats que propose la forêt d'arbres décisionnels. Nous avons également observé que Web Dice permettait de construire de meilleurs modèles comparé à Web PMI, réduisant le taux d'erreur en moyenne de 4.7 %. Nous avons enfin remarqué que les meilleurs résultats sont obtenus en combinant le score

6. Voir https://cran.r-project.org/web/packages/nnet/nnet.pdf

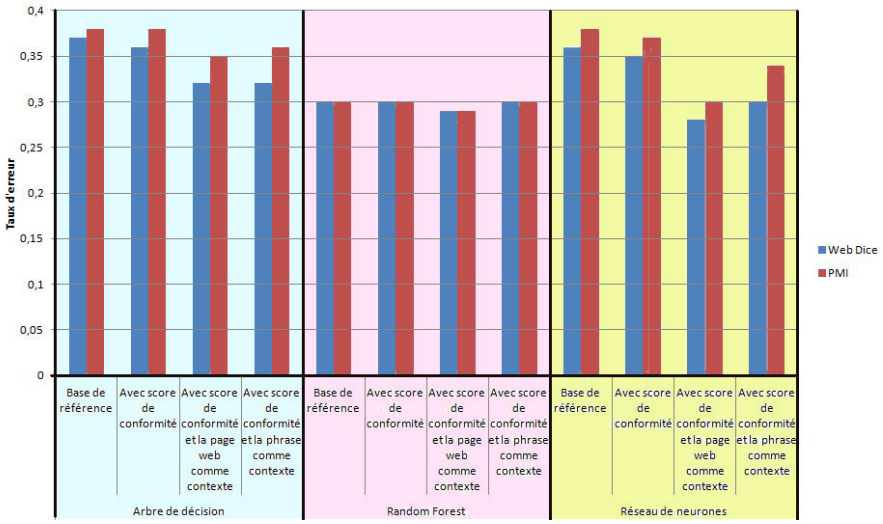


FIG. 2 – *Diagramme représentant le taux d'erreur en fonction de la configuration des modèles et de la formule pour calculer les mesures de proximité sémantique.*

de conformité et la page web comme contrainte pour le contexte. En effet, avec cette configuration, nous avons identifié correctement 72 % des relations sémantiques testées. Ce score montait à 82 % en supprimant les relations « mixtes » et 90 % en supprimant également les relations causales.

A travers cette expérience, nous avons observé que combiner le score de conformité, le contexte de la page web et la formule de Web Dice permet une relativement bonne identification de la classe sémantique des relations sémantiques.

7 Discussion

Comme observé par Costa et al. (Costa et al., 2011), certaines relations sémantiques appartiennent à plusieurs classes sémantiques, comme par exemple :

— « device » qui est à la fois un hyperonyme et un méronyme de « computer » ;

— « music » qui est à la fois une relation causale et un méronyme de « emotion ».

Notre méthodologie ne permet d'associer qu'une unique classe sémantique par relation sémantique. Travailler dans un environnement multi-classes pour l'apprentissage permettrait ainsi d'améliorer le taux de relations correctement identifiées.

8 Conclusion

Notre étude a confirmé l'efficacité des mesures de proximité sémantique pour l'identification de la classe de relations sémantiques. Nous avons correctement identifié 72 % de re-

lations sémantiques dans un environnement fortement bruité. Nous avons prouvé qu'imposer une contrainte sur le contexte, au niveau du document, permet l'amélioration des résultats. Le score de « conformité » a lui aussi un impact positif sur la qualité des résultats. Notre dernière observation porte sur les formules des mesures de proximité sémantique, où nous avons démontré que Web Dice permet une meilleure identification que Web PMI.

De futures améliorations sont possibles, notamment en modifiant le calcul du score de conformité. En effet, ce score est fondé sur une méthode relativement simple de comparaison qui gagnerait à être enrichie par les nombreux travaux qui portent sur l'acquisition de relations sémantiques.

Références

Berland, M. et E. Charniak (1999). Finding parts in very large corpora. *Proceedings of ACL-1999*, 81–106.

Bollegala, D., Y. Matsuo, et M. Ishizuka (2007). Measuring semantic similarity between words us-ing web search engines. *Proceedings of 16th International Conference on World Wide Web*, 81–106.

Bollegala, D., Y. Matsuo, et M. Ishizuka (2010). A web search engine-based approach to measure semantic similarity between words. *Knowledge and Data Engineering, IEEE Transactions 23*.

Breiman, L. (2001). Random forests. *Mach. Learn. 45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, et R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Caraballo, S. (1999). Automatic acquisition of a hypernym labeled noun hierarchy from text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 120–126.

Cederberg, S. et D. Widdows (2003). Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. *CONLL '03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 4*, 111–118.

Costa, H., H. Oliveira, et P. Gomes (2011). Using the web to validate lexico-semantic relations. *EPIA 2011 4*, 597–609.

De Oliveira, P. (2009). *Probabilistic Reasoning in the Semantic Web using Markov Logic*. Ph. D. thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering.

Girju, R. et D. Moldovan (2002). Text mining for causal relations. *FLAIRS 2002*, 360–364.

Girju, R. et D. Moldovan (2003). Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 1–8.

Girju, R. et D. Moldovan (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 83–135.

Gledson, A. et J. Keane (2008). Using web-search results to measure word-group similarity. *22nd International Conference on Computational Linguistics*, 281–288.

Haribabu, M., H. Devaraj, et R. Kumar (2014). A web search engine application to compare semantic equivalence between two words. *International Journal of Computer Science and Information Technologies 5*, 6572–6577.

Harris, Z. (1954). Distributional structure. *Word 10*, 146–162.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*.

Islam, A., E. Milios, et V. Keselj (2012). Comparing word relatedness measures based on google n-grams. *Proceedings of COLING*, 81–106.

Kozareva, Z., E. Riloff, et E. Hovy (2008). Semantic class learning from the web with hyponym pattern linkage graphs. *In Proceedings of the Association for Computational Linguistics*.

Lippmann, R. (1987). An introduction to computing with neural nets. *ASSP Magazine, IEEE 4*(2), 4–22.
English

Oshiro, T., P. Perez, et J. Baranauskas (2012). How many trees in a random forest? In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition*, Volume 7376 of *Lecture Notes in Computer Science*, pp. 154–168. Springer Berlin Heidelberg.

Pantel, P. et D. Ravichandran (2004). Automatically labeling semantic classes.

Sahami, M. et T. Heilman (2006). A web-based kernel function for measuring the similarity of short text snippets. *Proceedings of 15th International World Wide Web Conference*.

Tufféry, S. (2012). *Data Mining et Statistique Décisionelle. Quatrième édition*. Paris : TECHNIP.

Summary

Semantic relations are the core of a growing number of knowledge-intensive systems. The need to discover such relations led to significant advances in their automatic acquisition. However, their automatic validation remains an up-to-date challenge. In this paper, we present a web-based method enabling the automatic identification of the class of a semantic relation. Our hypothesis is that it is possible to determine the proximity between semantic relations and semantic classes by adding in the similarity measures expression, syntactic patterns indicative of semantic classes. Using these patterns as entry features for a learning algorithms, we are able to identify semantic relations. Our experiment resulted in successfully labelising 72 % of semantic relations divided in 3 semantic classes : hypernym, meronym and causal, in a rich semantic environment with 1 mixed class composed of antonyms, synonyms and co-hyponyms.

Index

A

Akdag Herman 1
Amy de la Breteque Estelle 1

B

Blansché Alexandre 43
Bonneau Anne 13
Boudjeloud-Assala Lydia 43

C

Cadot Martine 13
Chauchat Jean-Hugues 51
Compagno Dario 43
Conan-Guez Brieuc 43
Cornuejols Antoine 31

D

Dray Cedric 25

G

Girard Yann 62
Gély Alain 43

M

Mellouli Nédra 1
Mercier Arnaud 43
Mohamed Aichetou 1

O

Orobinska Olena 51

P

Pasquier Louis-Victor 31
Pepin Lambert 62
Petit Jean 73
Pinson Suzanne 31

R

Raynaut William 25
Risch Jean-Charles 73

S

Saurel Pierre 62
Sharonova Natalia 51
Soule-Dupuy Chantal 25
Stoichita Victor 1

V

Valet Philippe 25
Valles-Parlangeau Nathalie 25