TALN, Text-Mining et ontologie pour la maintenance de panneaux photovoltaïques

Sami Bouhouche

TALN, Text-Mining et ontologie pour la maintenance de panneaux photovoltaïques

BOUHOUCHE
Sami

Sous la direction de Claude PONTON

Entreprise : EDF

UFR LLASIC
Département I3L

Mémoire de master 2 mention Sciences du langage - 20 crédits

Parcours : Industries de la langue

Année universitaire 2018-2019

Remerciements

Je tiens à remercier toutes les personnes qui m'ont apporté leur aide durant mon stage et mon parcours universitaire.
Je remercie mes tuteurs en entreprise, S. M. et B. R. qui m'ont guidé tout au long du stage et se sont montrés disponibles et à l'écoute. Ils m'ont transmis de nombreux savoir-faire et surtout des valeurs qui me seront utiles durant tout le reste de mon parcours professionnel.
Je remercie Claude Ponton, mon tuteur universitaire pour ses conseils tout au long du stage ainsi que pour sa pédagogie durant mon parcours universitaire. Je le remercie également d'avoir fait partie des personnes qui m'ont donné goût à l'informatique et ce dès 2014 alors que j'étais en Licence de Langues étrangères Appliquées.
Je remercie également C. D. ainsi que l'ensemble des membres de l'équipe P16 pour l'accueil qu'ils m'ont réservé et leur bienveillance à mon égard.
J'exprime ma gratitude pour Mr Georges Antoniadis, responsable du master industries de la langue de l'université Grenoble Alpes ainsi que l'ensemble du corps enseignant qui m'ont, de par leur pédagogie et leur bienveillance, permis de me familiariser avec un domaine qui m'était encore inconnu il y'a 2 ans de cela.
Je remercie mes parents pour leur soutien inconditionnel durant le stage et bien au-delà. Papa, Mama aucun mot ne peut quantifier la gratitude et l'amour que je ressens pour vous.
Je remercie mes frères pour leur soutien indéfectible, leur présence dans les moments de joie mais surtout dans les moments plus difficile.
Je tiens à remercier les amis qui m'ont tendu la main et ont été un soutien de qualité tout au long de mon périple universitaire.
Je remercie également les stagiaires avec qui j'ai partagé le bureau V001 pendant la durée de mon stage. Votre sympathie a permis de faire de ce bureau l'un des plus joyeux du site.
Je remercie mes camarades de promotion qui m'ont accompagné durant les 2 ans de Master IdL.
Enfin, je remercie toutes les personnes que j'ai oublié de remercier et qui ne m'en tiendront pas rigueur.

Déclaration anti-plagiat

Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : BOUHOUCHE ……

PRENOM :Sami……

DATE : 07 Septembre 2019……

Sommaire

Remerciements 3
Sommaire 6
Table des illustrations 6
Liste des Tableaux 7
Introduction 8
Partie 1 - Contexte du stage 10
CHAPITRE 1. PRESENTATION DE EDF 11
1. LE GROUPE EDF 11
2. LA R&D D'EDF 12
3. EDF RENOUVELABLES 13
CHAPITRE 2. LE PROJET TEXT-MINING EOLIEN 14
Partie 2 - Fouille de texte : Etat de l'art 16
CHAPITRE 3. CONCEPTS DE BASES 17
CHAPITRE 4. PRINCIPES D'UN PROGRAMME DE FOUILLES DE TEXTE 19
1. APPROCHES SYMBOLIQUES 19
2. APPROCHES CONNEXIONNISTES 20
CHAPITRE 5. APPLICATIONS/TACHES DE LA FOUILLE DE TEXTE 21
1. LA RECHERCHE D'INFORMATION 21
2. LA CLASSIFICATION 21
3. L'EXTRACTION D'INFORMATION 21
4. L'ANNOTATION 21
CHAPITRE 6. L'OUTIL TEXT-MINING PREEXISTANT POUR L'EOLIEN 23
1. FONCTIONNEMENT DE L'APPLICATION 23
2. L'ONTOLOGIE 24
Partie 3 - Travail réalisé 26
CHAPITRE 7. ANALYSE ET STRUCTURATION DES DONNEES 27
CHAPITRE 8. ENRICHISSEMENT DE L'ONTOLOGIE 28
CHAPITRE 9. RESULTATS 32
1. DETECTION D'ELEMENTS PV 34
2. DETECTION D'ETATS 34
CHAPITRE 10. PERSPECTIVES ET EXPERIENCE PERSONNELLE 35
1. PERSPECTIVES 35
2. EXPERIENCE PERSONNELLE 36
Conclusion 37
Bibliographie 38
Glossaires 39

Table des illustrations

Figure 1 : répartition de la production électrique d'EDF en fonction des sources 11
Figure 2 : entités d'EDF SA 12
Figure 3 Approches symboliques 19
Figure 4: approche par apprentissage 20
Figure 5: Objectif de l'application 23

Figure 6 : Aperçu de la répartition des classes dans l'ontologie 25
Figure 7 Format d'un fichier d'entrée 27
Figure 8 Exemple d'objet non repéré 30
Figure 9 : Représentation de « Shelter » dans l'ontologie 32
Figure 10 : Détection de composants PV exemple 1 34
Figure 11 : Détection de composants PV exemple 2 34
Figure 12 Aperçu de la détection des états 34

Liste des Tableaux

Tableau 1 : Nombre d'éléments ajoutés à l'ontologie 31
Tableau 2 : résultats obtenus 33

Introduction

A l'heure où quasiment chaque action dans le monde réel se traduit par la création de données numériques, l'exploitation de celles-ci est devenue un enjeu majeur pour les entreprises et les institutions. La quantité de données numériques sur Terre est évaluée à 33 zetaoctects[1] - soit 33 milliards de giga-octets - et est amenée à croitre exponentiellement durant les années à venir.
Les données sont aujourd'hui présentes à tous les niveaux de fonctionnement des entreprises et des institutions. Les activités commerciales et techniques ainsi que les activités administratives, de support et de contrôle d'une entité en génèrent énormément. Ces données constituent la représentation numérique des activités d'une entité. Ainsi une compréhension fine de ses données à travers un travail d'analyse et d'interprétation permet de tirer de précieux enseignements quant au fonctionnement de l'entité et d'identifier des pistes d'amélioration d'optimisation. La quantité de données produites par une entreprise ou une institution rend la réalisation manuelle d'une telle tâche inenvisageable.
Plusieurs domaines se penchent sur les problématiques de traitement des données. Parmi eux un axe d'étude du traitement automatique du langage naturel se penche plus spécifiquement sur les problématiques de traitement de données de type textuelles: l'extraction d'information à partir de données. Cette branche du TAL consiste à traiter des données textuelles afin d'en synthétiser le contenu ou d'en tirer une information en particulier.
Ce mémoire de fin d'étude, réalisé à l'issue d'un master Science du langage parcours industries de la langue abordera donc les problématiques d'extraction d'information à partir d'une grande quantité de données. Le cas d'étude pratique sur lequel s'appuieront les éléments théoriques du présent mémoire est celui de l'optimisation d'un outil d'analyse textuelle pour la maintenance de panneaux photovoltaïques sur lequel je me suis impliqué durant 6 mois dans le cadre de mon stage de fin d'études effectué au sein du département PRISME de la R&D d'EDF. EDF, comme bon nombre d'entreprises, s'intéresse à l'optimisation de ses processus en tirant parti de l'analyse des données accumulées au cours des dernières décennies. C'est dans cette optique que s'inscrit le projet text-mining Eolien lancé en 2017 par le département PRISME et mis en œuvre par S. M. et B. R. qui m'ont encadré au cours de ce stage. Au cours de ce projet un outil se basant sur des règles et une ontologie a été mis en œuvre afin d'analyser des données issues de la maintenance de parcs éoliens pour en extraire des informations sur les actions réalisées et les composants affectés. Mon stage s'inscrit dans lacontinuité de ce projet et consiste à développer l'outil mis en œuvre afin qu'ils prennent en compte des composants, des actions ainsi que des états, dégradations et défauts spécifiques au domaine du photovoltaïque et ainsi l'adapter à l'analyse d'opérations de maintenance effectuées sur des parcs solaires.
La problématique régissant ce travail est la suivante : **Comment le traitement automatique des langues peut optimiser le processus de conception, d'exploitation et de maintenance d'installation photovoltaïques ?**
Ce mémoire s'articule de la manière suivante : tout d'abord, une présentation du contexte du stage et des motivations du projet, ensuite est abordé l'aspect théorique des techniques d'extraction d'informations et de text-mining afin de situer ce travail par rapport aux travaux antécédents, puis le fonctionnement de l'application et des améliorations mises en place ainsi que les résultats obtenus au cours du stage sont détaillés, enfin des pistes d'améliorations sont évoquées avant la conclusion de ce travail.

[1] « Quantité de données crées dans le monde », Tristan Gaudiot, https://fr.statista.com/infographie/17793/quantite-de-donnees-numeriques-creees-dans-le-monde/ , 24 Avril 2019 (consulté le 04/08/2019)

Partie 1

-

Contexte du stage

Chapitre 1. Présentation de EDF

1. Le Groupe EDF

Le sigle EDF renvoie au groupe électricité de France. Il s'agit du premier acteur de l'électricité en France ainsi qu'en d'Europe. Le groupe compte 165790 employés en 2018 [2] répartis entre EDF SA et ses différentes filiales.
Le champ d'activités du groupe couvre tout ce qui concerne l'électricité, de la production à la distribution en passant par des activités commerciales telle que le négoce d'énergie. Une division technique sépare les différentes activités d'EDF, on distingue :

Les activités Amont :

Les activités amont désignent tout ce qui concerne la production d'électricité.
La production électrique s'appuie sur un mix énergétique qui repose sur l'exploitation de différentes ressources :

- Production nucléaire
- Production Thermique
- Production à partir d'énergie renouvelables :
  - Hydraulique
  - Energie solaire
  - Eolien

La production électrique selon les différentes sources d'énergie se répartit comme suit :

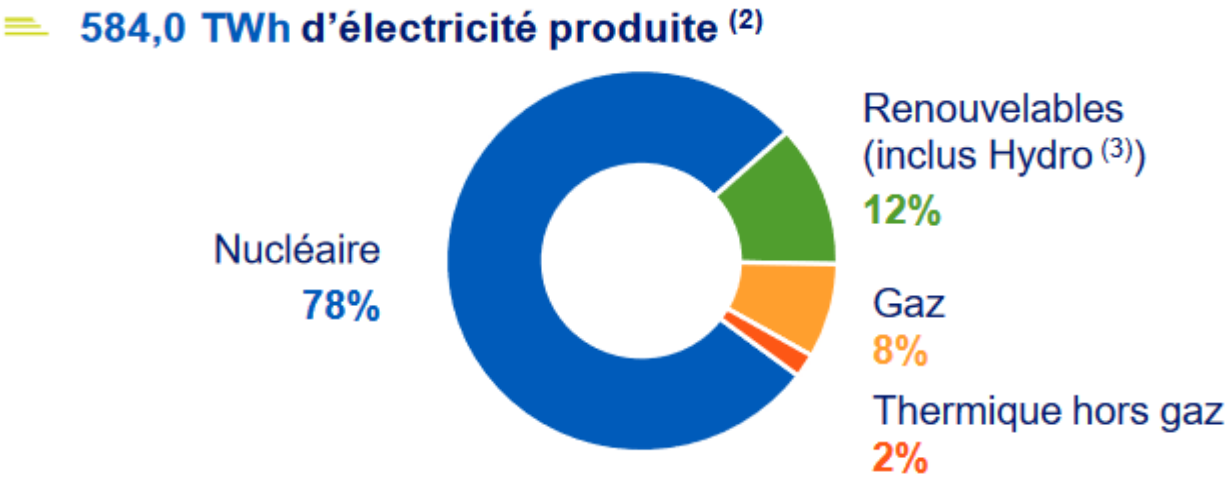


Figure 1 : Répartition de la production électrique d'EDF en fonction des sources [3]

[2] « EDF en Bref » Site officiel d'EDF https://www.edf.fr/groupe-edf/qui-sommes-nous/edf-en-bref (consulté le 02/08/2019)
[3] « Faits et chiffres 2018 », Document EDF, 2018 https://www.edf.fr/sites/default/files/contrib/groupe-edf/espaces-dedies/espace-finance-fr/informations-financieres/publications-financieres/faits-et-chiffres/facts-and-figures-2018-fr.pdf (consulté le 02/08/2019)

Les activités Avale :
Elles regroupent toutes les activités concernant le transport et la distribution d'électricité. Le transport est géré par RTE (détenu à 50.1% par EDF), Enedis (filiale 100% EDF, ex-ERDF) s'occupe de la distribution sur 95% du territoire français, les 5% restants sont gérés par les distributeurs non-nationalisés.

2. La R&D d'EDF

EDF SA s'articule autour d'entités spécialisées dans différents types d'activités :

Figure 2 : Entités d'EDF SA

La R&D d'EDF est rattaché à la Direction Innovation Responsabilité d'Entreprise Stratégie (DIRES).
La R&D regroupe environ 2100 salariés répartis sur 9 centres dont 3 en France : le site des Renardières, celui de Saclay ainsi que celui de Chatou dans lequel j'ai évolué durant ces 6 derniers mois. Ses principaux objectifs sont de contribuer à l'amélioration de la performance des unités opérationnelles du Groupe et de repérer les pistes de croissance à moyen et à long termes.
Le site de Chatou se subdivise en 3 départements de recherche : le LNHE (Laboratoire National d'Hydraulique et d'environnement), MFEE (Mécanique des Fluides, énergies, Environnement) et PRISME (Performance, Risques Industriels, Surveillance pour la Maintenance et l'Exploitation)
C'est dans ce contexte que le département PRISME a identifié les procédés de traitement automatique des langues comme potentiellement intéressants afin d'exploiter au mieux les

données de maintenance et a initié les projets de text-mining pour la maintenance. Il existe différentes données de maintenance :

1. Celles issues d'appareils de monitoring
2. Celles concernant la gestion des stocks et les mouvements de pièces (sorties/entrées)
3. Des rapports d'exploitation
4. Des rapports et comptes rendus de maintenance

Celles-ci sont stockées depuis 2012 pour le PV et l'éolien, un traitement adéquat permettrait d'en tirer de précieuses informations.

3. EDF Renouvelables

Dans le cadre de ce projet, le client est EDF Renouvelables. Il s'agit de la branche du groupe EDF qui s'occupe de la conception, l'exploitation et la maintenance de parcs électriques dont la production est basée sur des sources d'énergie renouvelables.

C'est plus précisément le département Technologies et Recherches pour l'Efficacité Énergétique (TREE) qui est pilote du projet R&D dans lequel s'inscrit ce travail. TREE est basé sur le site des Renardières et assure le support de R&D à EDF Renouvelables pour le développement et l'exploitation de l'énergie solaire photovoltaïque.

Chapitre 2. Le projet text-mining Eolien

Le projet text-mining Eolien est le premier d'une lignée de projets « text-mining » initiée par le département PRISME d'EDF en 2017. L'application fut enrichie en 2018 et est actuellement en cours d'industrialisation.
Pour comprendre l'intérêt de ce projet, il est important de saisir le fonctionnement du processus de maintenance à EDF Renouvelables.
On distingue deux types de maintenances :

- **La maintenance préventive** : Maintenance planifiée à échéance régulière, réalisée même lorsqu'il n'y a aucune panne constatée.
- **La maintenance corrective** : Maintenance pour pallier un dysfonctionnement. Elle se déroule de la manière suivante :
  - Le dysfonctionnement est généralement détecté par les systèmes de monitoring (surveillance, contrôle à distance) une alerte est émise.
  - L'alerte parvient à un opérateur qui se trouve dans un centre de conduite
  - L'opérateur tente de résoudre le problème à distance, s'il ne peut résoudre le problème à distance.
  - L'opérateur émet une demande d'intervention via un avis de service.
  - L'avis de service parvient à un responsable de zone.
  - S'il valide le besoin d'intervention et qu'il ne peut résoudre le problème à distance, le responsable de zone crée un ordre de travail (OT) afin qu'une intervention soit réalisée.
  - Un technicien intervient, si le problème est résolu à l'issue de l'intervention l'OT est alors clôturé, s'il ne parvient à résoudre le problème, un responsable de zone sollicite de nouvelles interventions.
  - A l'issu de chaque intervention des comptes rendus d'OT sont créés et stockés dans une base SAP (outil de gestion d'opérations industrielles), ils se présentent sous la forme de document contenant des données structurées (dates, nom de l'intervenant, poste d'intervention etc..) ainsi qu'un champ de texte libre que les techniciens rédigent à la fin de l'intervention.
  - L'équipe Support Terrain métier enquête sur les causes possibles de la panne, et cela peut passer par l'analyse de comptes rendus concernant des pannes similaires ou de l'historique d'une machine afin de repérer les actions précédemment menées et d'en déduire les causes possibles de la panne.

Cette dernière étape de la maintenance corrective peut s'avérer particulièrement fastidieuse et chronophage pour l'expert métier chargé de réaliser l'enquête de support. C'est donc à ce niveau que l'outil text-mining répond à un besoin métier.
La faisabilité de cette approche et le degré satisfaisant de précision et de qualité des résultats ont été établis à travers une démonstration où un expert métier avait dressé une liste d'éléments intéressants (action de maintenance, composant concerné par l'action) à partir de comptes rendus de maintenance d'une machine spécifique sur une période donnée afin de faire le point sur la situation après plusieurs interventions infructueuses. Les mêmes comptes rendus étaient analysés par l'application de text-mining. L'expert a identifié 19 évènements importants ; parmi ces 19 évènements 18 ont été reconnus au-moins partiellement par l'application. L'application a, de plus identifié une douzaine d'évènements supplémentaires jugés potentiellement intéressants par l'expert métier. Cette démonstration a mis en exergue l'intérêt d'une telle application et a encouragé EDF Renouvelables à poursuivre sa commandite de ces activités R&D text-mining.
Avant d'entrer dans les détails techniques du fonctionnement de l'application et des améliorations apportées à celle-ci, il est primordial de revenir sur des éléments théoriques et de dresser un panorama des approches du TAL pour l'extraction d'information et la fouille de texte.

Partie 2

-

Fouille de texte : Etat de l'art

La fouille de texte ou l'extraction de connaissance à partir de textes (ECT) plus communément désignée par l'anglicisme « text-mining » est à la croisée de plusieurs domaines. La première apparition du terme « ECT », du moins de son équivalent en anglais « knowledge discovery from text», date de 1995, et fut employé/utilisé par Feldman.[4]
Dans la littérature, il n'existe pas de consensus quant au positionnement théorique de cette discipline. Cependant il y'a une convergence vers le fait que la fouille de texte est étroitement liée à la fouille de données. Des travaux datant de la fin des année 90 tels que ceux de Tan (1999) situent la fouille de texte en tant que prolongement/spécialisation de la fouille de données (data mining) ou de l'extraction de connaissance à partir de base de données (ECBD)[5]. Ce positionnement s'est ensuite vue conforté par de nombreux travaux datant des années 2000[6] [7] [8]. Ainsi, ce travail se positionnera dans le sillon tracé par ces travaux.
La première partie de ce chapitre définit des éléments basiques nécessaires à la compréhension de ce travail, par la suite les deux principales approches de la fouille de texte sont présentées et puis, au terme de ce chapitre, la définition de « fouille de texte » adoptée dans ce travail est livrée.

Chapitre 3. Concepts de bases

- **Une donnée**

La donnée est un signal brut, non interprété. Les données peuvent être issues de capteur, d'environnement comme le web ou stockées dans des bases de données.
Elles peuvent se présenter sous 3 formes différentes :
Données non-structurées : Données dont la structure ne présente pas de régularité, par exemple du texte brut/libre.
Données semi-structurées : Données peu structurées mais contenant néanmoins une certaine régularité, les documents au format XML ou html en sont un bon exemple.
Données structurées : Données tabulaires, tableau Excel ou base de donnée.

- **Une information**

L'information est une donnée ou un ensemble de données qui a été interprétée. Par exemple, une valeur issue d'un pyranomètre (appareil de mesure de la puissance du rayonnement solaire) placé sur un panneau solaire présentée à un individu qui ne sait l'interpréter est une donnée, par contre un expert du domaine est en mesure d'interpréter cette donnée et de juger, par exemple si l'exposition d'un panneau photovoltaïque au rayon solaire a été satisfaisante ou non. L'interprétation de la donnée par l'expert en fait une information.

- **Une connaissance**

La connaissance est un ensemble de données et d'informations assimilé et utilisé pour assumer une tâche ou créer une nouvelle information. La connaissance est souvent caractérisée par deux

[4]" Knowledge Discovery in Textual Databases (KDT)" Ronen Feldman and Ido Dagan, 1995
[5] "Text Mining: The state of the art and the challenges", Ah-Hwee Tan , 1999
[6] "A Brief Survey of Text Mining" Andreas Hotho, May 2005
[7] "What Is Text Mining?" Marti Hearst, october 2003
[8] "A Survey of Text Mining Techniques and Applications" Vishal Gupta, 2009

éléments : la finalité car la connaissance est mise en œuvre pour atteindre un objectif et sa capacité générative puisqu'elle permet la création de nouvelles informations.

- **Une ontologie**

Le terme « ontologie » trouve son origine dans la « métaphysique » une branche de la philosophie et renvoi à l'étude de l'être. L'ontologie décline de l'approche d'Aristote pour qui la métaphysique est « la science de l'être en tant qu'être ». Autrement dit l'ontologie est la science qui étudie les êtres.

Le domaine de l'intelligence artificielle s'est ensuite emparé des ontologies et leur usage s'est démocratisé/répandu dans les systèmes à base de connaissances et s'est étendu aux problématiques d'extraction d'information. Ainsi, en informatique et en science de l'information, une ontologie est l'ensemble structuré des termes et concepts représentant le sens et les relations entre les éléments d'un domaine de connaissances.

Il existe différents types d'ontologies :

- **Les ontologies globales** (dite de « haut-niveau ») : présentent un haut niveau d'abstraction et de généralité. Elles ont pour but une utilisation générale (ex : WordNet).
- **Les ontologies de domaine, ou dédiées à une tâche plus spécifique** : sont limitées à la représentation des concepts d'un domaine en particulier (géographie, électricité, aviation etc.)
- **Les ontologies d'application :** offrent le plus fin niveau de spécificité, c'est-à-dire qu'elles sont créées pour résoudre un problème en particulier et sont dédiées à un champ d'application précis à l'intérieur d'un domaine. Par exemple, l'ensemble des composants et des défauts d'un parc solaire constituent une ontologie d'application qui spécifie les concepts généraux pouvant provenir d'une ontologie du domaine électrique général.

Chapitre 4. Principes d'un programme de fouilles de texte

Dans la littérature, on retrouve deux grandes approches pour aborder les problématiques de fouille de texte. Ces approches sont issues des 2 grands courants de pensée des domaines de l'intelligence artificielle et des sciences cognitives, il s'agit des approches symboliques et des approches connexionnistes. Il s'agit de deux vision du processus d'apprentissage : l'une tend à percevoir l'apprentissage comme un ensemble de règles et de symboles assimilés que l'on applique pour résoudre un problème donné, par exemple si un individu a appris la signification de l'alphabet d'un langage et assimilé les règles de lecture de celui-ci, alors il pourra lire n'importe quelle suite de caractères, il s'agit des approches dites « symboliques ».
Dans l'autre vision, un individu finit par assimiler des connaissances en étant confronté à des exemples sans pour autant connaitre la théorie derrière. Par exemple les premiers mots d'un nourrisson sont des mots auxquelles il a été confronté et qui apparaissent fréquemment dans son environnement, il finira par les prononcer en imitant/reproduisant ce qu'il perçoit, sans pour autant avoir compris les mécanismes de phonétiques articulatoires intervenant dans la prononciation de ce mot. Il s'agit des approches « connexionnistes ».

1. Approches symboliques

Les systèmes basés sur des approches symboliques sont des programmes réalisés par des experts ayant une bonne connaissance du domaine, du type de données et des résultats désirés (de l'objectif à atteindre par l'application). Leur fonctionnement repose sur un ensemble de ressources dont l'utilisation est détaillée par des règles. Dans le cadre du TAL ces ressources peuvent prendre la forme de lexique, dictionnaire ou dans notre cas, d'une ontologie.
Exemple de système utilisant des approches symboliques : automates, grammaire générative etc.

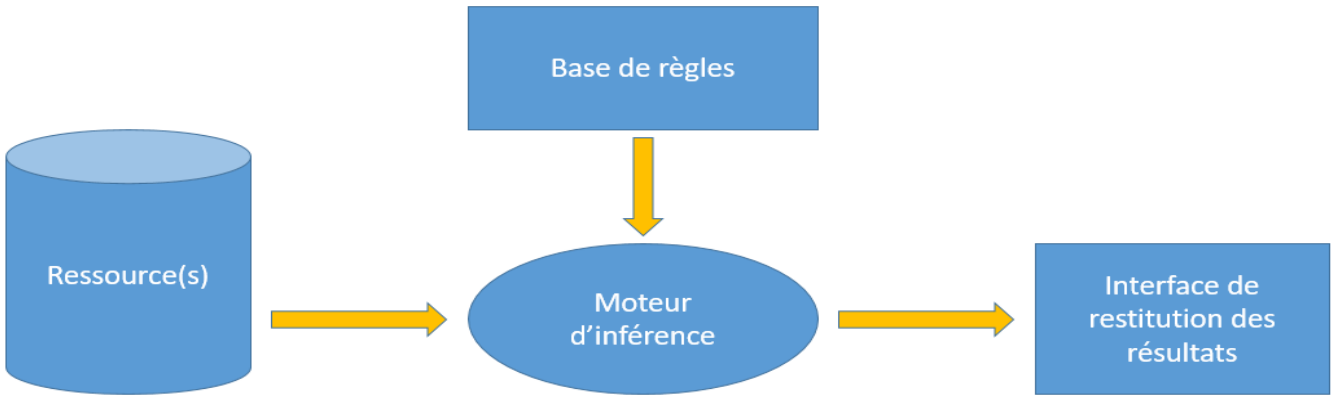


Figure 3 Approches symboliques

2. *Approches connexionnistes*

Les systèmes utilisent des algorithmes s'appuyant sur des méthodes statistiques.

Les programmes se basant sur des approches connexionnistes reposent sur des corpus annotés ou non et le résultat désiré.

Les programmes reposant sur ces approches sont assez rapides à implémenter. Ces programmes sont efficaces pour de grande quantité de données, cependant leur bon fonctionnement est étroitement lié aux données qu'il prend en entrée, ainsi avant de les appliquer une étape de préparation des données est indispensable.

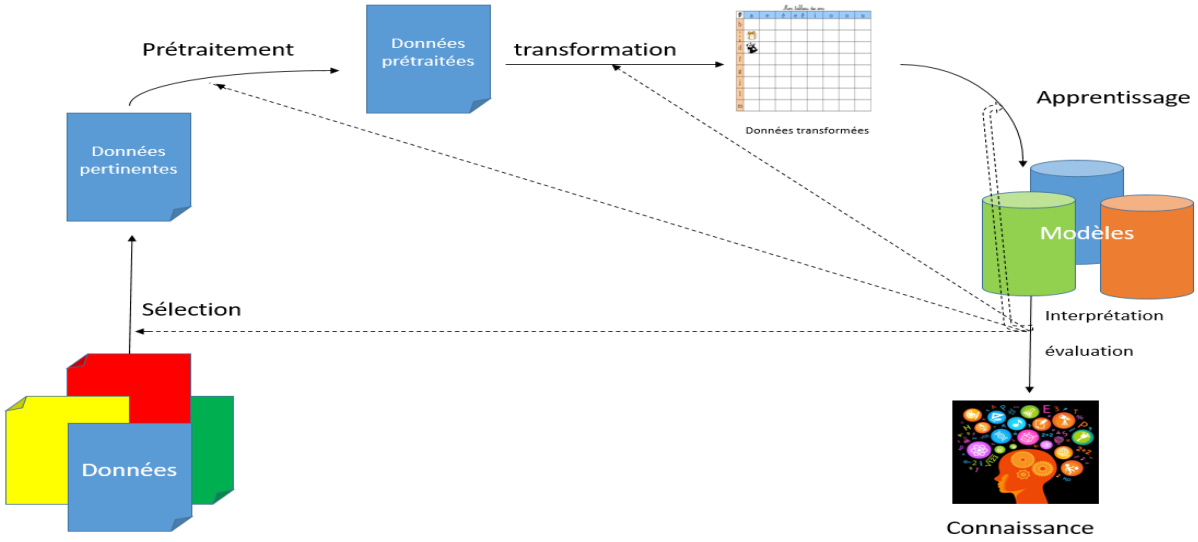


Figure 4: Approche par apprentissage

L'apprentissage automatique est utilisé dans de nombreux domaines : la détection de spam, la prédiction de prix, recommandation de produits, détection de fraude, conduite autonome, reconnaissance d'objets etc.

Chapitre 5. Applications/Tâches de la fouille de texte

1. La recherche d'information

La recherche d'information (RI) est un processus de sélection d'ensemble de documents pertinents, c'est-à-dire, ceux qui répondent le plus à une requête saisie par un utilisateur. Dans certains systèmes de RI, il existe différentes méthodes de suivi et d'analyse du comportement de l'utilisateur afin de retourner des documents selon son profil. Ces méthodes se basent sur des algorithmes d'analyse textuelles des profils pour assurer plus de satisfaction.

2. La classification

Classification : Cette tâche permet de regrouper différents documents/éléments dans un certain nombre de classes. La classification se fait selon un ou plusieurs critères de similarité.

3. L'extraction d'information

L'extraction d'information (EI) consiste à extraire des informations significatives à partir d'un ensemble de textes. Dans cette application, il y'a un besoin des ingénieurs du domaine qui permettent de valider les concepts du domaine et leurs relations selon leurs expertises. Ainsi cette technique est utilisée, non seulement, pour extraire des concepts bien précis à partir d'un ensemble de documents, mais surtout pour sélectionner les relations entre ces concepts-là. Généralement, les informations ainsi extraites sont stockées dans une base pour être réutilisées dans un but bien précis tel que la prise de décision.

4. L'annotation

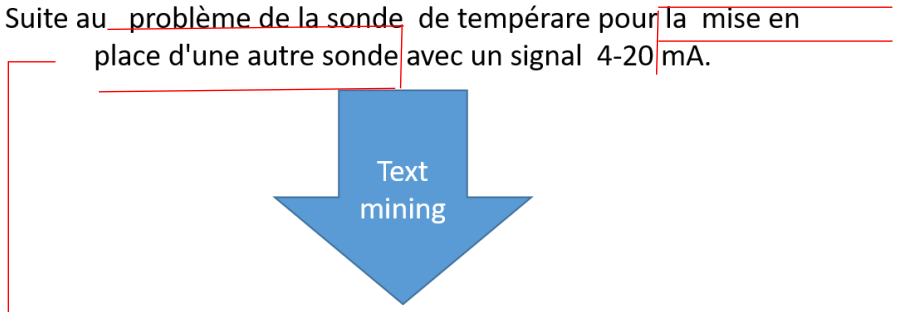
L'annotation consiste à associer des « étiquettes » à des données ou à des portions de texte. La longueur de la portion annotée est variable, dans le cadre du TAL les textes sont généralement découpé en tokens (mot, ponctuation).

L’utilisation que nous faisons ici de la fouille de texte consiste à rechercher et à extraire de l’information à partir de données textuelles puis à la restituer de manière structurée afin de construire de la connaissance.
Le projet Text-mining Eolien mis en place par le groupe P16 du département PRISME est un système à base de règle prenant en entrée des données semi-structurées (xml). La principale ressource sur laquelle il s’appuie est une ontologie d’application. Les principales tâches effectuées par celles-ci sont l’extraction d’information ainsi que l’annotation.
Le chapitre suivant détaillera les aspects techniques du fonctionnement de celle-ci.

Chapitre 6. L'outil text-mining préexistant pour l'éolien

1. *Fonctionnement de l'application*

L'application a été réalisée sous GATE (General Architecture for Text Engineering, une plateforme de développement d'applications permettant d'implémenter différents modules de traitement automatique des langues.
L'application s'appuie sur un ensemble de module de la plateforme GATE constituant une chaine de traitement (architecture dite en pipeline). Elle prend en entrée textes, qui correspondent aux comptes rendus de maintenance que l'on veut analyser. Ces comptes rendus font l'objet d'un traitement préalable puis structurés au format XML. Les données d'entrée contiennent des méta-données (technicien intervenant, dates et heure d'intervention etc …) ainsi qu'un champ de texte libre que les techniciens ont rédigé à la suite de leur intervention. C'est ce champ-là qui nous intéresse particulièrement et qui fera l'objet d'une analyse textuelle. Le but de cette analyse textuelle est de repérer des actions de maintenance, les composants concernés par ces actions ainsi que des constats sur l'état physique et fonctionnel de composants.
Voici un exemple pour illustrer cela :



Action	Objet de l'action	Complément	Type d'action	Type d'objet	Texte complet original	Position dans le texte
constat	sonde	problème de la sonde de tempérare pour la	inspection	capteur	Suite au problème de la sonde de tempérare pour la mise en place d'une autre sonde avec un signal 4-20 mA.	129424
installation	sonde | signal	mise en place d'une autre sonde avec un signal 4	nouveau	capteur	Suite au ||| problème de la sonde ||| de tempérare pour la ||| mise en place d'une autre sonde avec un signal ||| 4-20 mA. |||	129466

Figure 5: Objectif de l'application

La chaine opère en 5 étapes :

1. Récolte et sélection des données :
2. Nettoyage et structuration des données.
3. Fouille de texte.
4. Mise en forme des résultats.
5. Exploitation des résultats et stockage.

Les différents modules GATE utilisée dans l'application sont :

- *Document Reset PR* : réinitialise le document et supprime les annotations créées lors d'une exécution antérieure.

- *GATE Unicode Tokenizer* : découpe les textes en *tokens* (mots, nombres, ponctuation…).
- *ANNIE Sentence Splitter* : découpe le texte en phrases.
- *Generic Tagger* : utilise l'outil *TreeTagger* pour *lemmatiser*, c'est-à-dire identifier le lemme des termes rencontrés et pour étiqueter ceux-ci en fonction de leur nature (POS tagging)
- *Flexible Gazetteer* : utilisent les libellés définis dans l'ontologie pour associer aux termes ou groupes de termes correspondant à ces libellés les concepts qu'ils identifient dans l'ontologie.
- *JAPE Transducer* : ensemble de règles Jape. JAPE est le langage de règles implémenté dans GATE. Les parties « gauches » des règles décrivent des motifs et conditions sur les annotations créées par l'application GATE sur le texte traité ; la partie droite contient du code Java permettant de consulter, compléter, modifier, créer des annotations. Ce code peut en particulier accéder à l'ontologie du domaine utilisée comme ressource. Ces règles traitent les annotations identifiées par les modules précédents et explicitent la manière d'utiliser les éléments de l'ontologie.

Un point essentiel du fonctionnement de l'application est qu'elle s'appuie sur une ontologie en guise de ressource.

2. L'ontologie

L'ontologie mise en place est une ontologie d'application, elle regroupe les différents composants structurels, mécaniques, électriques et électroniques susceptibles d'être trouvés dans un parc éolien, des états et des défauts propres à ses composants, ainsi que tous les éléments (mêmes secondaires) qui ont pu être identifiés à travers l'analyse des comptes rendus de maintenance. Celle-ci a été créée via Protégé, un outil permettant de créer et de manipuler des ontologies au format rdfxml ou owl.
Les objets sont répartis en classe de la manière suivante :

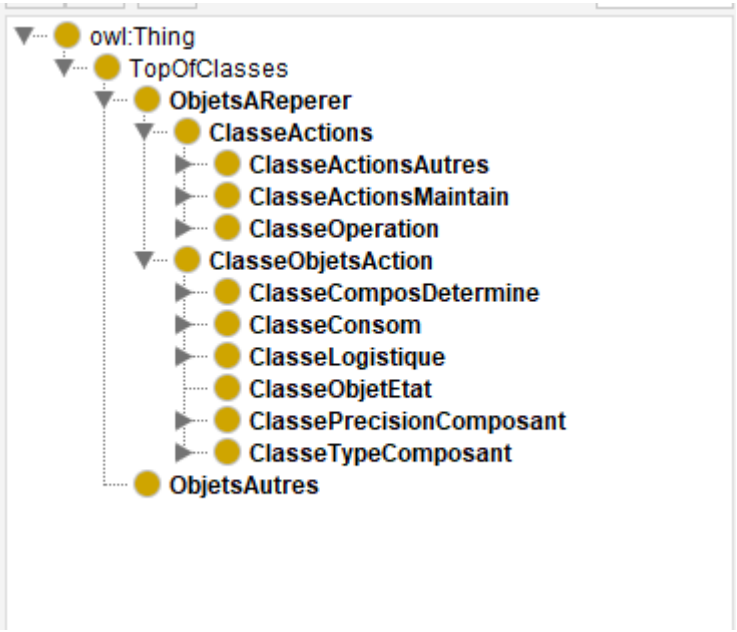


Figure 6 : Aperçu de la répartition des classes dans l'ontologie

Dans la figure 6 les classes ne sont pas entièrement déroulées, au total il y'a 817 instances (individus) réparties entre 71 classes. La principale division est celle séparant les actions et les objets. Les objets sont placés dans des sous-classes selon leur fonction (électrique, hydraulique etc.). Les actions sont organisées selon leur type (action de maintenance, autres actions etc.). L'ontologie étant applicative, sa raison d'être est de pouvoir être utilisée par les règles JAPE afin de repérer des éléments pertinents dans le texte, ainsi la structure de l'ontologie n'a pas pour but de représenter la réalité scientifique ou théorique du domaine de l'éolien mais de représenter un découpage du domaine qui soit « parlant » pour les spécialistes qui voudront analyser les informations extraites. C'est par exemple ce besoin vis-à-vis des spécialistes qui conduit à séparer les actions de « remplacement » des actions « d'entretien » car ils veulent pouvoir les examiner différemment.

Ainsi l'objectif est d'adapter l'application text mining au domaine de la maintenance de parcs solaires photovoltaïques.

Partie 3

-

Travail réalisé

Chapitre 7. Analyse et structuration des données

L'application prend en entrée des données au format xml, un exemple de document avant analyse :

Figure 7 Format d'un fichier d'entrée

Pour parvenir à ce format, il est nécessaire de rassembler, organiser et structurer les données. Les données à disposition sont des extractions issues de différents outils de gestion utilisés au sein d'EDF :

- Les comptes rendus d'ordre de travail
- Le journal du centre de surveillance d'exploitation
- La référence et le type des parcs
- La liste des pièces sorties pour chaque ordre de travail
- Les codes pannes des différentes machines : les codes pannes sont des alertes émises par les systèmes de monitoring lorsqu'une anomalie ou un dysfonctionnement est détecté.

Les données sont transmises par EDF Renouvelables. Celles-ci sont des extractions d'outils de gestion interne (le progiciel de gestion intégré SAP, l'outil de surveillance OCC tool par exemple), elles contiennent des données concernant les parcs solaires et éoliens et nous parviennent sous forme de fichiers MHTML, CSV et Excel. Les données sont tout d'abord converties si besoin puis importées dans *Access* (un système de gestion de base de données du *pack Microsoft Office* pour Windows). Cet import permet d'effectuer quelques étapes préliminaires (jointures, regroupements ou séparation de champs, conversion de caractères spéciaux…) grâce aux fonctionnalités d'Access avant de passer les données dans l'application GATE.
Cette partie du travail relève de la fouille de données, dans un premier temps, l'aspect textuel des données a peu d'incidence sur le traitement. Voici les différentes tâches réalisées pour mettre en forme les données :

- Unification
- Séparation PV/Eolien : Comme les données englobent le PV et l'Eolien, il était nécessaire d'isoler les données PV afin de pouvoir observer les textes qu'ils contiennent et repérer des composants et états propres aux PV. Cela a été réalisé en deux étapes : une jointure entre les parcs indiqués dans les comptes rendus d'OT et d'un fichier décrivant la nature des différents parcs et une sélection des enregistrements relatifs aux parcs désirés.
- Récupération des postes techniques :
- Normalisation des entêtes
- Mise sous forme concise et exploitable des dates et heures associées aux comptes rendus
- Jointure avec les pièces sorties

Chapitre 8. Enrichissement de l'ontologie

L'objectif de cet enrichissement est d'inclure dans l'ontologie les concepts pertinents pour identifier les informations recherchées dans les textes de comptes rendus relatifs au photovoltaïque, typiquement les actions de maintenance spécifiques, les composants, paramètres des parcs PV mais également les états possibles et dégradations de ces composants.

Identification des termes :
L'objectif est d'identifier les termes pouvant apparaitre dans les textes et faisant référence aux concepts constitutifs des informations cherchées et donc utiles à capitaliser dans l'ontologie.
La recherche de termes pertinents s'est faite à travers plusieurs méthodes.
La première consiste à parcourir les données et à relever manuellement les termes intéressants.
Une seconde approche a permis de repérer des termes. Elle s'appuie sur l'hypothèse selon laquelle les termes recherchés sont des termes techniques peu susceptibles de se retrouver dans un lexique « classique » (généraliste). Elle a été mise en œuvre via le logiciel Nooj (un outil

permettant de faire de l'analyse sémantique) et consiste à extraire les termes qui ne sont pas « reconnus » par Nooj. En effet Nooj permet de faire des extractions de lexique, or pour ce faire il s'appuie sur un lexique contenant des lemmes (forme « canonique » des mots) auxquelles sont associés des règles de flexion (ensemble des modifications susceptibles d'être apportées à un nom ou un verbe pour en exprimer différents aspects : genre, nombre, personne etc.). Ainsi, en utilisant la fonction qui permet d'afficher les mots inconnus de Nooj, on se retrouve avec une liste de 6000 entrées. Enfin après avoir nettoyé cette liste en éliminant les mots à 1 caractère, les ponctuations et suites de ponctuations ainsi que les éléments mal orthographiés, une analyse manuelle est faite. A l'issue de cette analyse, un script python a été mis en place afin de faciliter l'étape de sélection. Ce script permet d'obtenir un fichier Excel contenant les termes retenus, leur nombre d'occurrence ainsi que leur contexte d'apparition. Un champ vide permet d'émettre une suggestion sur les modalités d'ajout des termes dans l'ontologie. La figure ci-dessous montre un exemple de fichier obtenu avec ce script

terme	nombre d'occurence	contexte d'apparition	suggestion
pjb	280	entre les 2 onduleurs le jour de l'inspection. Vérification des tous les câbles entre PJB et SJB, SJB et stand = RAS Mesures à refaire lorsque les valeurs seront plus importantes (il est possible que le défaut soit lié avec un problème de connexion) Ordre 5200460 :	
dp10	71	Réparation de la cloture DP10 (refixation des socles maintenant le fil) sur environ 1/4 de la centrale#### occ3: Remplacement des modules DP10 Contrôle intrusion clôture (2 poteaux successifs pourris ainsi que la végétation haute aux abords de la clôtures	ajouter à la classe électro/info
sc	62	remplacement carte electronique SC 20 CONT (ticket SMA 401€).#### occ2: Intervention SMA MEA : 8h35 MES : 10h10 Arrivée sur site la carte SC com était allumée mais la com était HS Remplacement carte SC com, mais au démarrage des aux	
stack	45	echange de deux cartes , le probleme disparu en remettent les carte a leur place , redemarrage de onduleur a 12h05 prevoir un remplacement preventir de la carte#### occ2: remplacement des cartes drivers stack Intervention réalisée par SMA voir AT	
poweris	36	occ2: Acquittement des défauts batteries POWERIS SH01 et SH02.#### occ3: Tension d'alimentation des détecteurs incendie insuffissante : +/- 4,5V DC pour un fonctionnement à +/- 12V DC. Contrôle Coffret POWERIS + TDBT (alimentation	classe élec
tsa	34	l'arret des départs parcs et du TSA ouverture du disjoncteur HT et commande des disjoncteur HTB et du sectionneur de ligne , fonctionnement conforme. contrôle du MICOM P922 à suivre#### occ2: Consignation du transformateur TSA (ouve#### occ3:	
zonning	21	electricien avec changement des MC4 KC#### occ2: Rescensement par zonning (zone 4)#### occ3: Zonning Zone 3 effectué#### occ4: Zonning secteur 4 Changement MC4 Sh5-I2-S21#### occ5: Zonning zone 3 et lever de reserve sur c#### occ6: lever des	Classe action
DG	19	alarmes du Pdl. Reset du coffret c13-100. remise à 1 du DG. Vérification avec la conduite suite à un défaut au réenclenchement. L'ACR n'avait pas arrêté la demande de découplage. Remise à 1 du DG et vérification avec la conduite. Onduleurs ok. Départ	libellé possible de disjoncteur
PN	19	SH08-I1-S19-G01 PN 4/6 NON CASSE/HS/SHUNT L31- PN 8/4 NON CASSE/HS/SHUNT PN 5/2 CASSE/EN PRODUCTION/REPARATION CONNECTEUR MC4 PIEUVRE#### occ2: BALLON	libellé possible pour panneau

Figure 8 : Termes candidats et contexte d'apparition

Approche s'appuyant sur les résultats de l'application text-mining avant enrichissement :
Un passage de nos données dans l'application d'analyse textuelle avant l'enrichissement de l'ontologie a permis de repérer quelques termes. En effet lorsqu'une action est repérée et qu'aucun objet n'y est rattaché, cela signifie peut signifier deux chose : soit l'objet n'est tout simplement pas mentionné, soit il n'a pas été détecté car celui-ci (du moins la manière dont il est mentionné) n'est pas présent dans l'ontologie. Cela se matérialise de la manière suivante :

Num. OT	Num. ligne	Parc	N° machine	Action (TM)	Composant (	Texte pris en compte (TM)	Texte origin	Position dan	Identifiant ('	Type de site
5202476	2.0132E+13		SH01	arrêter	???	Shelter tjours à l'arrêt	Shelter tjour	58096	260168	SOLAIRE

Figure 9 Exemple d'objet non repéré

Dans l'exemple de la figure 7, aucun objet n'est associé à l'action « arrêter » cependant une analyse rapide permet de voir qu'il s'agit du terme « shelter ». Celui-ci est alors ajouté à la liste des termes candidats pour l'enrichissement de l'ontologie. Quelques discussions et recherches à travers les données ont permis de comprendre que « shelter » renvoi à un ensemble de panneaux photovoltaïques, il sera donc ajouté à l'ontologie.

La combinaison de ces 3 approches a permis de constituer une liste de 400 termes intéressants.

- Ajout des termes :

L'ajout des termes s'est fait via l'application Protégé. Lorsqu'un terme renvoie à un concept qui n'est pas représenté par une instance dans l'ontologie, son ajout passe alors par la création d'une nouvelle instance. Lorsque celui-ci renvoi à un concept pour lequel il existe déjà une instance dans l'ontologie, il est alors ajouté en annotation de celle-ci. Cette activité d'ajout de termes et d'instances dans l'ontologie n'est pas toujours triviale et peut parfois demander quelques recherches complémentaires («qu'est-ce qu'on désigne par 'pyranomètre'?») voire d'interroger un spécialiste du domaine.

- Nombre d'éléments ajoutés :

Protégé fournit Quelques métriques pour les ontologies. Ci-dessous une comparaison des métriques avant et après l'enrichissement :

Ontologie originale

Metrics

Axiom	4191
Logical axiom count	909
Declaration axioms count	895
Class count	72
Object property count	1
Data property count	0
Individual count	819
Annotation Property count	4

Annotation axioms

AnnotationAssertion	2387
AnnotationPropertyDomain	0
AnnotationPropertyRangeOf	0

Nouvelle ontologie

Metrics

Axiom	4949
Logical axiom count	1057
Declaration axioms count	1043
Class count	74
Object property count	1
Data property count	0
Individual count	965
Annotation Property count	4

Annotation axioms

AnnotationAssertion	2849
AnnotationPropertyDomain	0
AnnotationPropertyRangeOf	0

Tableau 1 : Nombre d'éléments ajoutés à l'ontologie

Les classes (*Class count*) : seulement 2 classes ont été ajoutées. Ce faible nombre s'explique par le fait que la structure de l'ontologie n'a pas été modifiée. Celle-ci étant étroitement liée au règles JAPE et au fonctionnement général de l'application, une modification de celle-ci requiert un ajustement des règles et peut éventuellement impacter d'autres modules de la chaine de traitement. Les deux classes ajoutées sont celles contenant les états repérés (ClasseEtat, ClasseEtatParam).

Les instances (*individual count*): 146 nouvelles instances ont été ajoutées. Les instances correspondent aux différents composants et états.

Les annotations : Les annotations sont des propriétés associées aux instances, l'application s'appuie sur deux types d'annotations :

- ***LabelSynthèse* :** On y retrouve la manière la plus courante de désigner un composant. C'est une propriété unique, c'est-à-dire que chaque instance à un seul *labelSynthèse* associé.
- ***Libellepossibles* :** Ce sont les différentes manières de désigner un objet (synonyme, équivalent en anglais, abréviation, terme utilisé dans le jargon etc.). Cela permet de lier les formes textuelles retrouvées dans le texte aux éléments de l'ontologie qu'elles désignent. Une instance peut avoir plusieurs *LibellePossible*.

  462 annotations ont été ajoutées.

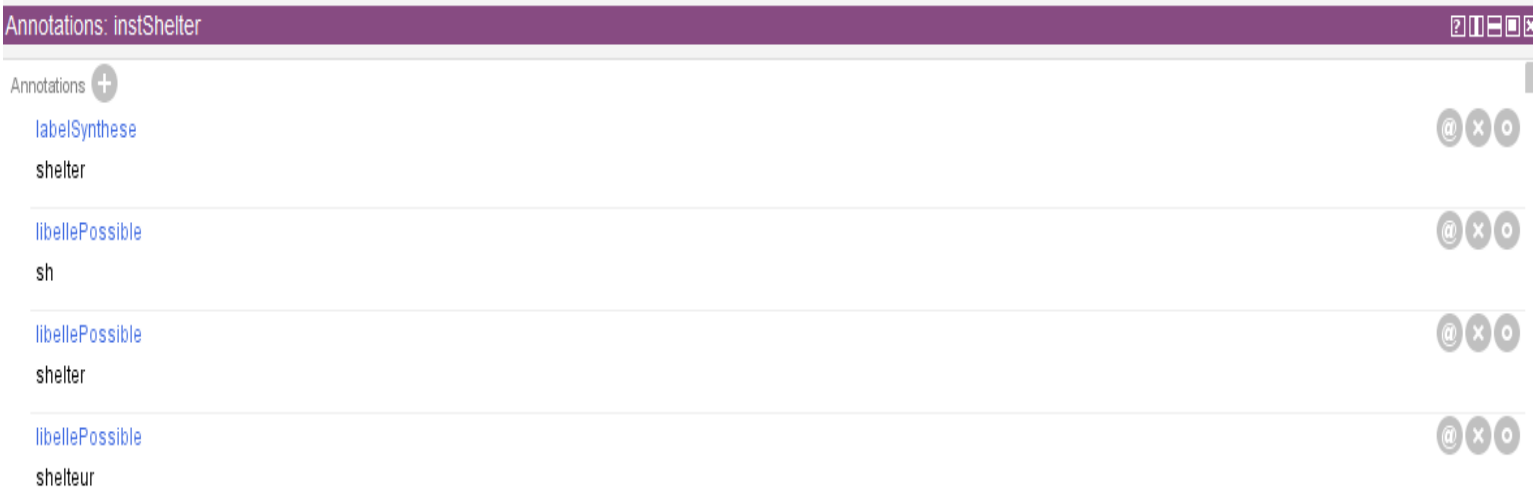


Figure 10 : Représentation de « Shelter » dans l'ontologie

A titre d'exemple, la figure 8 illustre la manière dont le terme « shelter » est représenté. Différentes annotations « libellePossible » lui sont associées ; parmi celles-ci on retrouve une abréviation (« sh ») , sa forme courante (« shelter ») ainsi qu'une manière alternative de l'orthographier (« shelteur ») qui a pu être observée à travers l'analyse des données.

Chapitre 9. Résultats

L'évaluation des résultats s'est faite en adoptant la même démarche que Stéphanie Nogueira qui a travaillé sur l'enrichissement de l'ontologie pour l'Eolien. La démarche consiste à comparer le nombre de d'opérations et d'objets repérés avant et après l'enrichissement. Sur les 55000 lignes de départ, 21833 contiennent du texte et sont exploitables.

	Avant l'enrichissement	**Après l'enrichissement**	**Comparaison**
Nombre d'enregistrements analysés (données d'entrées)	55000	55000	0
Nombre d'actions détectées	27726	30973	**+3247**
Nombre d'objets détectés	23960	27574	**+3614**
Nombre d'objets non repérés	3766	3399	**-367**
Nombre d'états repérés	Non traités en tant qu'état	3089	**+3089**
Pourcentage d'objets identifiés par rapports au nombre d'actions	86,42%	89,03%	**+2,61%**

Tableau 2 : résultats obtenus

Le pourcentage d'objets identifiés est un indicateur pénalisant car parmi les objets non identifiés, il y a ceux qui ne sont pas mentionnés dans le texte (et ne peuvent donc pas être reconnus par l'application) en plus de ceux qui sont mentionnés dans le texte mais que l'application ne reconnaît pas.
Les mesures « classiques » telles que la précision ou le rappel n'ont pu être calculées car le nombre total d'élément à repérer n'était pas connu.
Les résultats montrent la prise en compte d'actions et d'objets supplémentaires. Il est intéressant de les regarder de plus près tout en vérifiant s'il sont cohérents avec les deux axes principaux de ce travail, c'est-à-dire la détection de composants PV ainsi que la prise en compte des états, défauts et dysfonctionnements. Il faut noter qu'une partie des éléments non-repérés ne sont simplement pas mentionnés dans le texte des comptes rendus

1. *Détection d'éléments PV*

Afin d'avoir une idée précise sur la détection d'éléments spécifiques au domaine de l'énergie solaire, nous pouvons, dans le fichier Excel obtenu suite au traitement, filtrer sur des composants typiques du PV, par exemple le terme « shelter ».

ID	operation	ordre	Action	Objet de l'action	Type d'action	Texte lu
187645	10	5239230	test overtemperature	shelter	inspection	Test échauffement SH

Figure 12 : Détection de composants PV exemple 2

Les figures ci-dessus permettent de voir que les composants ajoutés à l'ontologie sont bel et bien pris en compte. Ils sont d'autant plus parlant car le terme « Shelter » est reconnu en tant qu'objet d'action de maintenance alors qu'il est désigné par deux formes différentes : sa forme classique (« shelter ») ainsi qu'une abréviation (« SH »).

2. *Détection d'états*

L'un des apports majeurs de la nouvelle version de l'application est une meilleure prise en compte des états et des défauts. Les états sont indiqués dans les résultats par la présence de la mention « état » dans le champ type action.

ID	operation	ordre	Action	Objet de l'action	Type d'action	Texte lu
150971	10	5230779	overtemperature	onduleur	etat	surchauffe de l'onduleur
161523	10	5233585	faible | isolement	onduleur	etat	faible isolement onduleur à surveiller
164917	10	5234555	à l'arrêt	onduleur	etat	Onduleur à l'arrêt
174415	10	5235774	en faute	onduleur	etat	onduleur en faute 9000 récurente
194555	10	5241115	bon fonctionnement	onduleur	etat	bon fonctionnement des 8 onduleurs

Figure 13 Aperçu de la détection des états

L'échantillon de la figure 13 présente un aperçu des états détectés. Il a été obtenu en filtrant les résultats obtenus sur les actions de type « état » et « onduleur » en objet. Le filtrage sur les onduleurs présente un grand intérêt car les onduleurs sont des composants essentiels dans une chaine de production électrique. Repérer les différents états de défaillance évoqués dans les comptes rendus et en identifier les causes peut permettre de prendre des mesures pour éviter l'apparition de ces états à l'avenir. Cela illustre bien le fait que les résultats peuvent constituer une aide à la prise de décision.
À contrario, l'analyse des états positifs ou du passage d'un état de défaillance à un état positif et l'identification des actions menées pour arriver à ces états peut mener à la capitalisation des bonnes pratiques.

Enfin, le dernier point relevé à travers l'analyse des résultats est qu'il est difficile d'évaluer la pertinence de ces résultats sans l'avis d'un expert métier.

Chapitre 10. Perspectives et expérience personnelle

1. Perspectives

La prochaine étape du projet est la restitution des résultats aux métiers afin qu'il puisse dans un premier temps les valider. Une fois validés, les informations contenues dans les résultats pourront être exploitées afin de créer des connaissances.
Des améliorations sont envisageables au niveau de la restitution des états et défauts. La détection des états et défauts est une fonctionnalité récente de l'application, qui certes fournit des résultats intéressants, mais doit cependant être développée d'avantage. Les améliorations doivent notamment porter sur la restitution car pour le moment les états repérés sont restitués sous le champ « action », ce qui peut créer une légère ambiguïté.
Les résultats, jusqu'à présent encourageants, peuvent déboucher sur une prise de conscience par tous les acteurs de la maintenance de l'intérêt de l'action text mining . On peut même aller plus loin en imaginant que cela puisse avoir un impact positif sur l'organisation et la gestion des données et ainsi avoir des données de mieux en mieux organisées et bien formées. Si tel est le cas, l'exploitation des données et l'analyse par l'application text mining en sera facilitée. On se retrouvera ainsi dans un cercle vertueux ou les informations extraites de données débouchent sur la création de connaissances ; ces connaissances seront capitalisées par la suite et pris en compte lors de l'analyse de nouvelles données.
Par ailleurs, une phase de déploiement doit être envisagée afin de pouvoir utiliser l'application dans un contexte industrielle ou des données sont créées en continus. Au vu de la sensibilité des données, il faudra être bien s'assurer, lors de cette phase, que la sécurité des données est garantie du début à la fin du processus.
On peut, outre les éléments cités ci-dessus, envisager de :

- Demander à un expert métier d'annoter des comptes rendus afin d'avoir une base de comparaison pour évaluer la pertinence des résultats obtenus

- il peut être intéressant de combiner la démarche proposée avec des approches par apprentissages (en se servant des annotations obtenus pour alimenter un programme d'apprentissage supervisés et réaliser des classifications automatiques pour les nouvelles données par exemple)
- Etendre l'application à d'autre domaine de l'énergie voir de la maintenance en général : Il suffit d'enrichir l'ontologie avec les éléments de ce domaine et celui-ci sera pris en compte par l'application.

2. Expérience personnelle

D'un point de vue personnel, le sujet du stage effectué s'inscrit dans le prolongement d'une lignée de problématiques vers lesquelles je me suis orienté au cours de mon cursus universitaire : la constitution de corpus textuels à partir de sources de données variées, puis l'extraction d'informations à partir de ces corpus.

Ce stage m'a permis de voir comment ces problématiques se matérialisent concrètement et sont gérées dans une grande entreprise telle qu'EDF et ainsi de mieux situer la place du traitement automatique des langues dans un processus industriel.

Au-delà des compétences relevant de ma formation, cette expérience m'a également permis d'enrichir mes connaissances du domaine du photovoltaïque et de l'électricité en général. Ces connaissances ont été acquises à travers l'analyse des données à disposition, des discussions avec nos clients et d'autres membres du département PRISME, mais aussi au travers de recherches personnelles.

Il aurait été intéressant de confronter ces connaissances à la réalité du terrain en se rendant sur un parc solaire, afin d'en observer le fonctionnement et dialoguer avec les différents acteurs de la maintenance pour cerner au mieux leur besoin.

Conclusion

Pour faire face au besoin d'optimisation des données d'EDF Renouvelables, nous avons exploré une solution d'analyse des textes des comptes rendus d'intervention fondée sur une approche basée sur un ensemble de règles ainsi qu'une ressource de type ontologie décrivant le domaine considéré.
Au cours de ce travail, une chaine de traitement de texte conçue pour le domaine de l'éolien a été reprise et adaptée pour la prise en compte d'éléments du domaine photovoltaïque. Cette adaptation s'est faite à travers l'enrichissement de la principale ressource sur laquelle s'appuie l'application : une ontologie.
Les résultats de départ montrent que les éléments d'installations solaires étaient assez bien pris en compte, cependant une amélioration a pu être observée grâce à cet enrichissement. L'enrichissement a également permis la détection des états.
Les résultats obtenus sont encourageants et laissent penser que la présence d'une brique TAL est tout à fait cohérente avec l'optimisation de données de maintenance.

Bibliographie

I. " Quantité de données crées dans le monde ", Tristan Gaudiot, https://fr.statista.com/infographie/17793/quantite-de-donnees-numeriques-creees-dans-le-monde/ , 24 Avril 2019 (consulté le 04/08/2019)

II. " EDF en Bref " Site officiel d'EDF https://www.edf.fr/groupe-edf/qui-sommes-nous/edf-en-bref (consulté le 02/08/2019)

III. " Faits et chiffres 2018 " , Document EDF, 2018 https://www.edf.fr/sites/default/files/contrib/groupe-edf/espaces-dedies/espace-finance-fr/informations-financieres/publications-financieres/faits-et-chiffres/facts-and-figures-2018-fr.pdf (consulté le 02/08/2019)

IV. " Knowledge Discovery in Textual Databases (KDT)" Ronen Feldman and Ido Dagan, 1995

V. "Text Mining: The state of the art and the challenges", Ah-Hwee Tan , 1999

VI. "A Brief Survey of Text Mining" Andreas Hotho, May 2005

VII. "What Is Text Mining?" Marti Hearst, october 2003

VIII. "A Survey of Text Mining Techniques and Applications" Vishal Gupta, 2009

IX. " Introduction à la fouille de textes" université de Paris 3 - Sorbonne Nouvelle, I. Tellier

Glossaires

CP : Code panne

CR : Compte Rendu

DIRES : Direction Innovation Responsabilité d'Entreprise Stratégie

EDF : Electricité De France

GATE: General Architecture for Text Engineering

Ontologie : Représentation de l'ensemble des termes et concepts d'un domaine.

OT : Ordre de Travail

PV : Photovoltaïque

PRISME : Performance, Risques Industriels, Surveillance pour la Maintenance et l'Exploitation

R&D : Recherche et Développement

RZ : Responsable Zone

TAL : Traitement automatique des langues

TREE : Technologies et Recherches pour l'Efficacité Énergétique

XML : eXtensible Markup Language

MOTS-CLÉS : TAL, Ontologies, Maintenance, Fouille de texte, Parcs solaires

RÉSUMÉ

L'abondance de données qui circulent dans le monde, conséquence de l'avènement du numérique, s'accompagne de questionnements quant aux usages qui peuvent être fait de ces données.
C'est ce besoin qui a conduit EDF à lancer des projets de recherches afin d'exploiter les données à sa disposition. Parmi ces projets, l'un se penche sur les données textuelles rendant compte des opérations de maintenance effectuées dans les parcs électriques solaires. Le travail réalisé dans ce mémoire a été accompli dans le cadre de ce projet. L'objectif est d'exploiter les données textuelles et d'en extraire des informations afin d'optimiser le processus de maintenance dans les parcs photovoltaïques.
La problématique est la suivante : Comment le traitement automatique des langues peut-il permettre d'optimiser le processus de conception, d'exploitation et de maintenance d'installation photovoltaïques ?
Pour répondre à cette problématique plusieurs actions ont été menées afin d'adapter une application d'analyse textuelle au domaine solaire. L'application repose sur une ontologie et des règles. Les principales actions décrites dans ce mémoire ont pour but de mettre en forme les données de départ et d'aboutir à l'enrichissement de l'ontologie afin que l'application soit compatible avec le traitement de données textuelles provenant de comptes rendus de maintenance de panneaux solaire.

KEYWORDS : NLP, Solar parks, Maintenance, Ontology, Text-Mining

ABSTRACT

The very large amount of circulating data, around the world is related to the rise of the digital. Many questions stand out as to the uses that can be made of these data.
As a resort EDF aims to launch research projects to make use of the data at his disposal. Among these projects, one looks at textual data. The analysis conducted in this thesis was established as part of this project. The aim is to use textual data and extract information to optimize the maintenance process in photovoltaic parks.
The issue is: How can natural language processing optimize the process of designing, operating and maintaining photovoltaic systems?
To answer this, several actions have been carried out in order to adapt a textual analysis application to the solar domain. The application lays on an on an ontology and rules. The main actions described in this thesis are intended to format the initial data in order to extend the ontology, so that the application is compatible with the processing of textual data, from maintenance reports of solar panels.