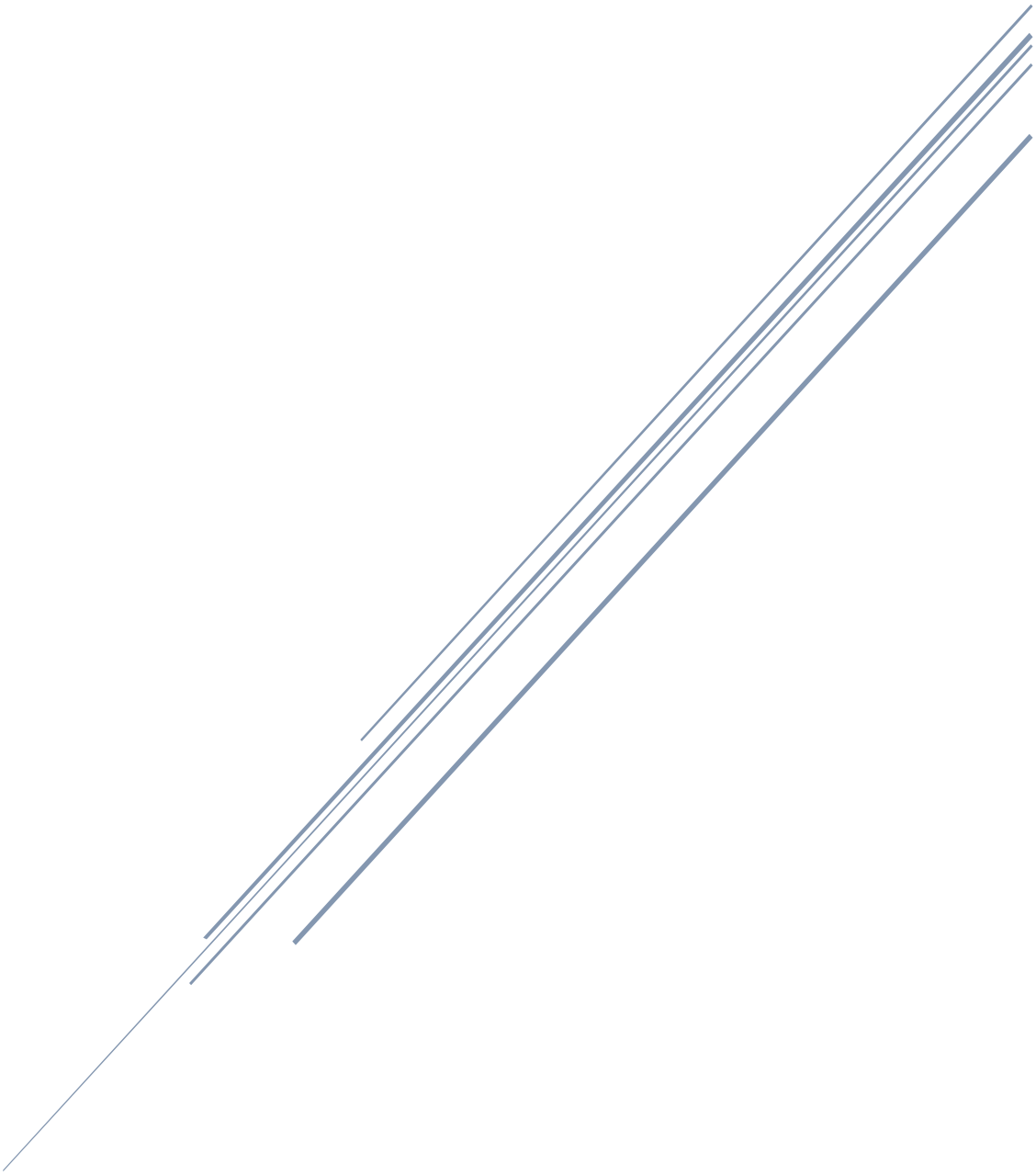
RAPPORT EISD

Bommart – Galmant – Madelaine – Ribet



Polytech Paris Sud – ET5 Informatique – 2018
T. Lavergne – S. Rosset

Table des matières

Cadre du Projet

Présentation du Sujet

Au tout départ, nous envisagions de créer un système de dialogue orienté vers la politique généraliste. Très vite, nous nous sommes rendu compte que le sujet était bien trop vaste et complexe pour ce projet.

Nous avons donc décidé de nous orienter vers les personnalités politiques, leur nom, leur prénom, leur date de naissance, leur famille, leurs partis, etc…

Après avoir analysé les différents corpus à notre disposition, nous avons décidé de préciser encore un peu plus le thème de notre système de dialogue pour le focaliser sur les personnalités politiques françaises.

Objectifs Initiaux

Les objectifs que nous nous sommes initialement fixés étaient, pour la plupart, maladroitement formulés et assez éloignés du but principal qu'un système de dialogue, qui est axé sur les personnalités politiques, devrait avoir. Ces objectifs étaient d'ailleurs plus proche de ceux d'un chatbot, créé dans le simple but de divertir l'utilisateur, que d'un réel assistant politique.

On peut tout de même garder quelques-uns de ces objectifs, à savoir :

Objectifs Initiaux	Statut
Indiquer clairement à l'utilisateur ce qu'il est possible de faire afin d'éviter toute frustration	
Permettre à l'utilisateur de changer de sujet quand il le veut sans altérer l'échange	
Prendre en compte les réponses utilisateur afin de s'auto corriger	
Reconnaître les négations	

Légende	
	Réalisé
	moyen
	Non fait

FIGURE 1 TABLEAU RECAPITULATIFS DES OBJECTIFS INITIAUX

Et c'est, guidé par ces objectifs, que nous nous sommes plongés dans ce projet.

Répartition du Travail

Nous avons alors séparé le projet en deux parties : une partie liée au Dialogue utilisateur et une autre partie liée à la Connaissance (partie gérant toute les informations mise à la disposition du système de dialogue). C'est donc autour de ces deux parties que nous avons basé toute l'architecture du projet ainsi que notre répartition des tâches : Cédrick et Manfred sur de la Partie Dialogue tandis qu'Hugo et Léo s'occupent de la Partie Connaissance.

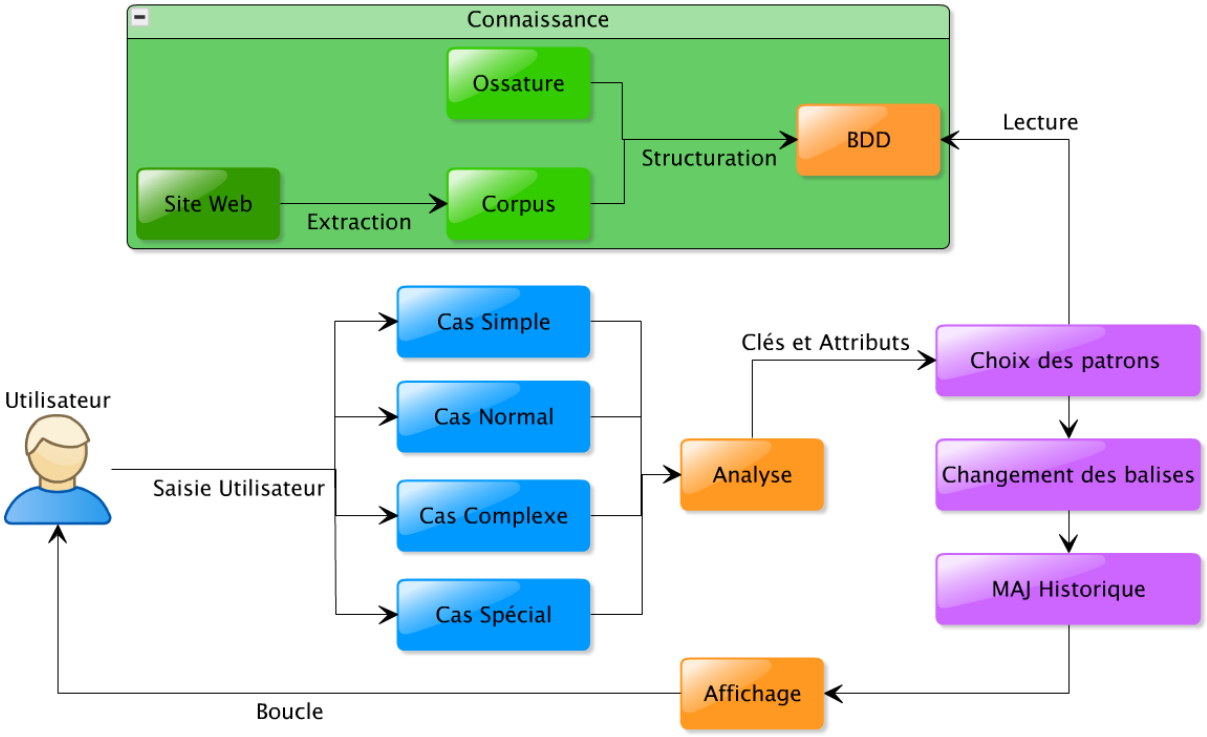


FIGURE 2 ARCHITECTURE DU PROJET

Partie Dialogue

Route vers la Généricité

Nous avons tout d'abord implémenté une version très basique du système de dialogue à savoir, l'équivalent d'un système de Question/Réponse sans analyse mais avec condition d'arrêt.

FIGURE 3 VERSION BASIQUE DU SYSTEME DE DIALOGUE

Une fois ceci fait, nous avons étoffé le système de dialogue en y ajoutant l'analyse de questions simples et la génération de quelques réponses en fonction des patterns reconnus. Puis nous avons étendu cette reconnaissance de patterns à quelques attributs de la base de données.

Mais, à chaque ajout de fonctionnalité dans le projet, ce dernier gagnait en complexité et devenait de plus en plus dépendant du sujet, si bien qu'il était devenu extrêmement compliqué d'ajouter quoi que ce soit au projet, sans perdre un temps considérable à comprendre les rouages de celui-ci.

Afin de solutionner ce problème, nous avons repris tout le projet en essayant de le rendre le plus générique et le plus modulaire que possible. Et c'est comme cela que nous nous sommes détaché du caractère spécifique du sujet pour tendre plutôt vers un système de dialogue presque indépendant du sujet donné.

Et c'est en faisant ceci que nous est venu l'idée de créé une grammaire afin décomposer et visualiser clairement toutes les questions susceptibles d'être posées et ainsi avoir une bonne analyse des questions. Cette grammaire nous a permis de regrouper plusieurs cas en un seul et de gagner en simplicité. Puis nous avons créé des modèles de réponses prêt à l'utilisation et contenant uniquement des balises qui seront substituées par la suite avec les données réelles d'un dialogue en cours.

Grammaire et Analyse des questions

Grammaire

Comme dit précédemment, nous avons créé une grammaire très sommaire permettant d'identifier le plus de questions possibles, et c'est de cette grammaire que découle toute

l'analyse de nos questions, allant du cas simple cas complexe en passant par le cas spécial. ainsi, cette grammaire est composée de **3 règles** (Élément, Question et Phrase) et de **3 opérateurs** (Négation, Produit et Somme), à savoir:

E ::= Clé | Attribut

| ! E | (P)

Q ::= Q * E

| E

P ::= P + Q

| Q

Ainsi, en se basant sur cette grammaire, nous pouvions reconnaître les cas suivants :

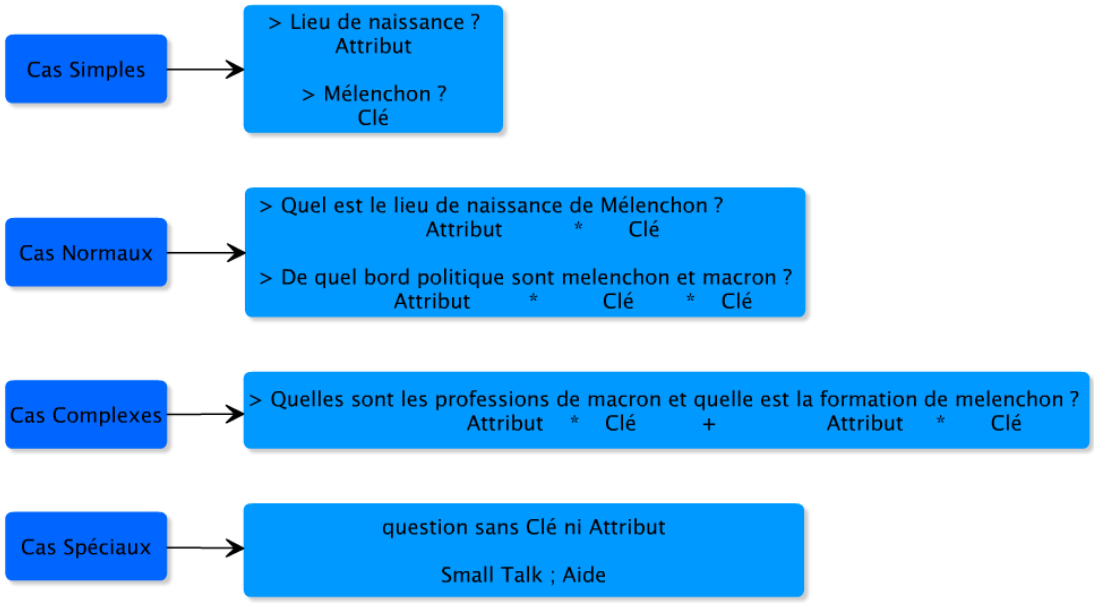


FIGURE 4 EXEMPLES DE QUESTION RECONNUES PAR LE SYSTEME DE DIALOGUE

Nous avons donc essayé de suivre le plus que possible cette grammaire mais certains cas se sont avérés difficiles à gérer, notamment la somme de n questions pour $n > 2$. On ne gère, à l'heure actuelle, que les cas où $n <= 2$ car, il est compliqué de délimiter clairement le commencement ainsi que la fin de n questions combinées en une seule.

Gestion des Clés

Dans un premier temps, l'accès aux politiciens se faisait via leur nom uniquement, mais cette méthode d'accès aux données soulevait le **problème de l'unicité des clés**. En effet, comment faire pour différencier 2 clés (ou plus) du même nom ? Nous avons alors généré une **clé unique**, par politicien, qui correspond à la fusion du prénom et du nom de celui-ci (en assumant que le cas où deux politiciens auraient la même appellation, soit négligeable). Ainsi, pour Jean-Luc Mélenchon, par exemple, on obtiendra comme clé "jean-luc_melenchon". Remarquons qu'il y a eu un travail de formatage de la clé afin de retirer les accents, les majuscules ainsi que les espaces.

Une fois cette tâche accomplie, il restait encore à reconnaître quand un politicien était mentionné dans une question et ce qui n'a pas été chose facile. En effet, il nous fallait une liste, exhaustive, de tous les noms présents dans notre base de données afin que tous ces politiciens soient accessibles à l'utilisateur. Nous avons donc stocké dans un fichier à la fois le nom et le prénom des politiciens en même temps que nous générions la base de données, puis nous l'avons nettoyé et ordonné par ordre décroissant de longueur afin d'obtenir le fichier final "*pnominal.txt*" contenant une seule instance de chaque nom ou prénom de politicien.

Le choix du tri dans l'ordre inverse de la longueur des éléments a été fait car nous nous heurtions à un problème : "Quelle est la date de naissance de Jean-Luc ?", si le nom "jean" précède le nom "Jean-Luc", alors la clé sera découpée en deux partie : la première correspondant au "jean" et la seconde à ce qui est reconnu avec le reste la clé initiale "Luc…", chose que l'on ne souhaite pas. Nous avons donc mis tous les noms de petite taille tout en bas du fichier.

Sur les 7469 politiciens présents dans notre base de données, on obtenait une liste de 23093 prénoms ou noms bruts, liste qu'on allait ensuite parcourir n fois durant un dialogue, avec n la somme des mots (plus petite entité délimitée par des espaces) présents dans toutes les questions posées. Cependant, grâce au prétraitement de cette liste de noms, on est passé de 23093 lignes, à 6819 en retirant toutes les répétitions et mots interdits. On a donc divisé par 3,4 le nombres de lignes à parcourir et par 3,2 la taille du fichier lui-même.

Grâce à ce fichier, nous pouvons maintenant accéder aux données d'un politicien via son *prénom* uniquement, son *nom* uniquement, ou son *prénom* suivi de son *nom* ! (Le cas où l'on voudrait accéder aux données d'un politicien via le *prénom* suivi du *nom* n'est volontairement pas géré car par convention, le prénom vient toujours avant le nom).

Dans certains cas, il peut y avoir ambiguïté sur le politicien auquel on doit associer la clé entrée par l'utilisateur, dans ces cas-là, le système passe en mode *script guidé*, affichant la liste des politiciens touchés par l'ambiguïté et permettant à l'utiliser de lever cette dernière. Cependant, ce passage en mode *script guidé* entraîne une coupure assez brutale dans le cours du dialogue, c'est pourquoi il reste encore un peu de travail à faire pour que la transition vers ce mode soit plus fluide.

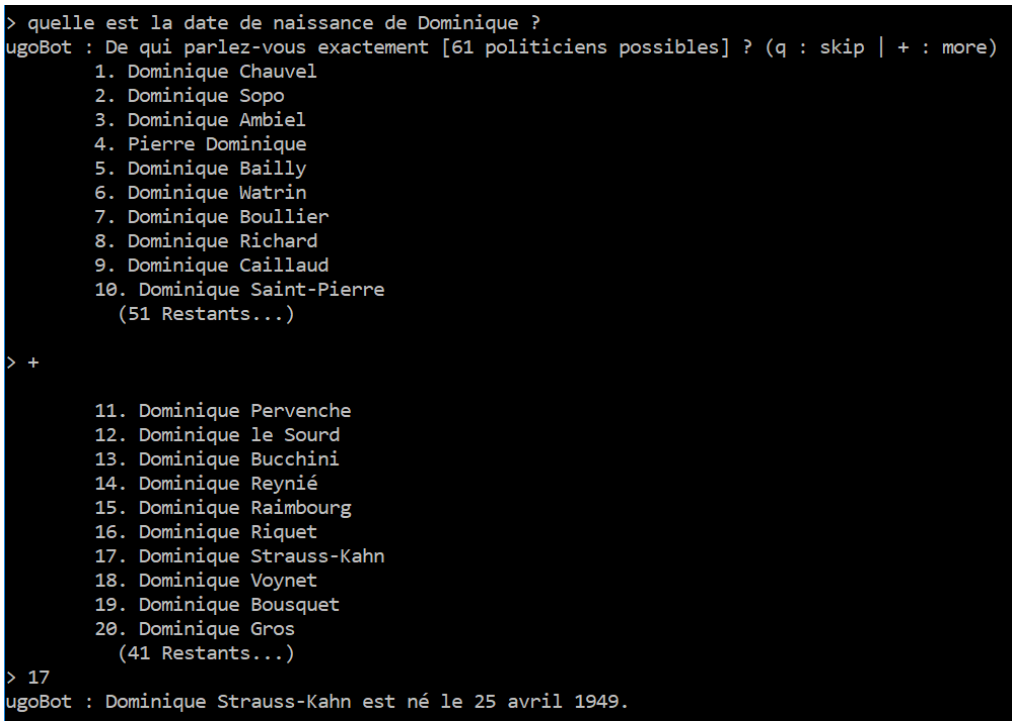


FIGURE 5 EXEMPLE DE PASSAGE EN MODE SCRIPT GUIDE APRES LA SAISIE D'UNE CLE AMBIGUË PAR L'UTILISATEUR

Clé Particulières

En plus des clés liées aux politiciens, nous avons également implémenté d'autres type de clés permettant d'interagir différemment avec le Système de Dialogue. En voici la liste :

Clé	Utilité
$help	Clé d'aide permettant d'indiquer à l'utilisateur ce qu'il est possible de faire avec ce système de dialogue afin d'éviter toute frustration
Historique	Clé permettant d'afficher l'historique de dialogue courant. Quatre informations y sont stockées : la question courante, la liste des clés et celle des attributs de la question et la réponse
Tutoiement	Clé permettant à l'utilisateur de poser des questions sur le système
Utilisateur	Clé permettant à l'utilisateur de poser des questions sur lui-même

Ces autres clés sont là uniquement pour étoffer et enrichir le système de dialogue, leur approfondissement ne pourra se faire que si le cœur du projet, à savoir la gestion des politiciens, est solidement construit. C'est pourquoi nous avons préféré laisser de côté ces pistes de réflexion, non pas que les idées nous manquaient, bien au contraire, mais parce que nous voulions avoir une accessibilité des attributs convenables ainsi qu'un système robuste.

Afin de développer le caractère divertissant du système de dialogue nous pensons enrichir l'environnement des clés de Tutoiement et de l'Utilisateur, notamment en finissant la personnification du système et en ajoutant du Small Talk. Ceci permettra d'ajouter un peu de liant et de vie aux divers dialogues, exemple d'échange à implémenter :

> Je ne t'aime plus.

ugoBot: moi non plus, j'étais très bien jusqu'à ce qu'on me dérange.

Gestion des Attributs

Pour répondre à notre besoin d'indépendance, nous nous sommes créé une base de données provisoire, "*databaseBeta.lua*" pendant que le groupe chargé de la Connaissance s'attelait à la génération de la base de données finale à partir du corpus. Nous avons donc travaillé pendant une très longue période indépendamment les uns des autres et, ce qui devait arriver arriva malheureusement, nous nous sommes retrouvés à implémenter des choses, qui étaient prévues au départ, mais qui se sont avérés compliquées à extraire du corpus et donc supprimées comme attributs principaux. Ainsi l'attribut primaire "Bord Politique" a été rendu accessible à l'utilisateur, dans la version Bêta, mais se retrouve finalement dépourvu de valeurs dans la version finale.

Mis à part ce petit soucis de communication, nous avons quand même pu créer un accès à tous les attributs primaires d'un politicien (**accessibilité de l'attribut**) via au moins une question utilisateur.

Attribut Primiare	Attribut Secondaire	Accessibilité
Nom	/	
Prénom	/	
Pronom (Il/Elle)	/	
Date de naissance	/	
Lieu de naissance	/	
Date de décès	/	
Formation	Diplôme/Titre	
	Date	
	Lieu	
Profession	Intitulé	
	date début	
	date fin	
Parti	Nom	
	Acronyme	
	Date d'adhésion	
	Date de départ	
	Particule	
Famille	Statut	
	Nom	
	Prénom	
	Métier	
Bord politique	Nom	
	Particule	

FIGURE 6 TABLEAU D'ACCESSIBILITE DES ATTRIBUTS D'UN POLITICIEN

Gestion des particules

Nous utilisons des particules afin de rendre les réponses du système plus authentiques :

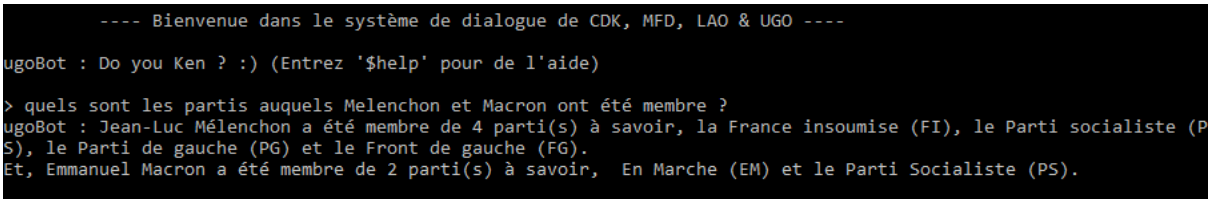


FIGURE 7 EXEMPLE DE REPONSE SYSTEME ENRICHI DE PARTICULES PRECISANT LE GENRE DES PARTIS

Utilisation de Pronoms

La langue française apprécie moyennement les répétitions hasardeuses, or notre système de dialogue est obligé de préciser le sujet dont il est question dans sa réponse afin de ne pas introduire de l'ambiguïté. Ceci engendre donc des répétitions inutiles du sujet, à savoir le prénom + le nom du politicien dont il est question. On a donc introduit une particule gérant le pronom à associer au politicien (Il ou Elle) afin d'effectuer des substitutions en cas de répétition avec le sujet et de tendre ainsi vers des réponses qu'un vrai assistant aurait dites.

La substitution s'opère cependant uniquement dans certains cas bien précis :

(Q : Question, K : Key, A : Attribute)

Il y aura substitution si la question se trouve dans l'un des contextes ci-dessous :

Q : K *(A + A) => K * A + Pronom * A

(Q1 : K * A) + (Q2 : K * A) => K * A + Pronom * A

Mais s'il n'y a pas succession directe entre deux clés identiques, rien ne se passe :

(Q1 : K1 * A) + (Q2 : (K2 + K1) * A) => pas de substitution

Afin de déterminer si la clé précédente est bien la même que la clé courante, on s'aide un peu de l'historique.

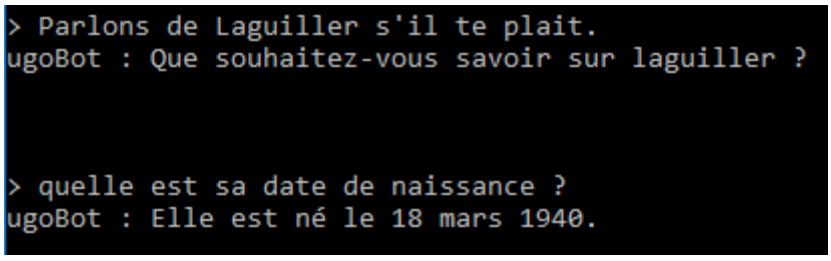


FIGURE 8 EXEMPLE DE SUBSTITUTION DU SUJET PAR SON PRONOM

Gestion des Attributs Secondaires

L'accessibilité des Attributs Secondaires (A.S.) se fait majoritairement par le biais d'un attribut primaire : l'Utilisateur voit le contenu d'un A.S. lors de l'affichage de la réponse à un A.P. qui lui est lié. L'accès direct à un A.S., c'est-à-dire en posant une question sur ce fameux attribut, n'est que partiellement fait. Nous souhaitons avoir une accessibilité des A.S. de 100%. Pour l'instant, nous estimons que 75% d'entre eux sont accessibles soit directement soit par le biais d'A.P.

Gestion des cas limites

Nous avons essayé de couvrir le plus de failles que possible, ainsi, voici les cas que sont gérés par le système :

Liste des cas
L'utilisateur rentre une chaine de caractère non reconnue par les patterns
L'utilisateur pose une question incomplète : phrase avec une clé ou un attribut uniquement
L'attribut ou la clé est reconnu mais il n'y a pas cette information dans la base de données

Modèle & Réponse aux questions

Modèle et Balises

Afin de bien gérer les réponses système nous avons mis en place plusieurs modèles de réponses toutes faites attendent simplement qu'on remplace leurs balises par des données réelles. L'implémentation de ces modèles nous a fait gagner énormément en clarté. En effet, en regroupant toutes les réponses systèmes dans un seul fichier, responsable de la gestion des modèles et des balises, nous a aidé à voir les relations entre les différents types de questions et de s'en inspirer.

Chaque réponse est tirée aléatoirement dans la liste du modèle sélectionné. Nous avons pu enrichir notre éventail de réponses en soumettant un questionnaire de réponses utilisateur à des proches, ce qui nous a permis de diversifier les types de réponses.

Gestion des Répétitions

Tout comme on ne souhaite pas répéter deux fois le sujet dans une même réponse, on ne veut pas non plus être amené à répondre strictement la même chose à l'utilisateur plusieurs fois de suite. Ainsi, pour les questions utilisateur qui sont en réalité plusieurs questions en une seule, on supprime toutes les réponses en double avant d'afficher la réponse finale à l'utilisateur.

Avec ceci de fait, on peut pousser la réflexion plus loin et commencer la chasse aux doublons, ainsi, si l'utilisateur me demande 2 fois la même chose, on ne veut pas lui afficher bêtement la même réponse mais lui montrer plutôt qu'on a vu qu'il y avait une répétition. Par exemple,

une réponse de ce genre “Vous voulez vraiment me faire répéter une seconde fois ?” serait une réponse à proposer dans ce cas.

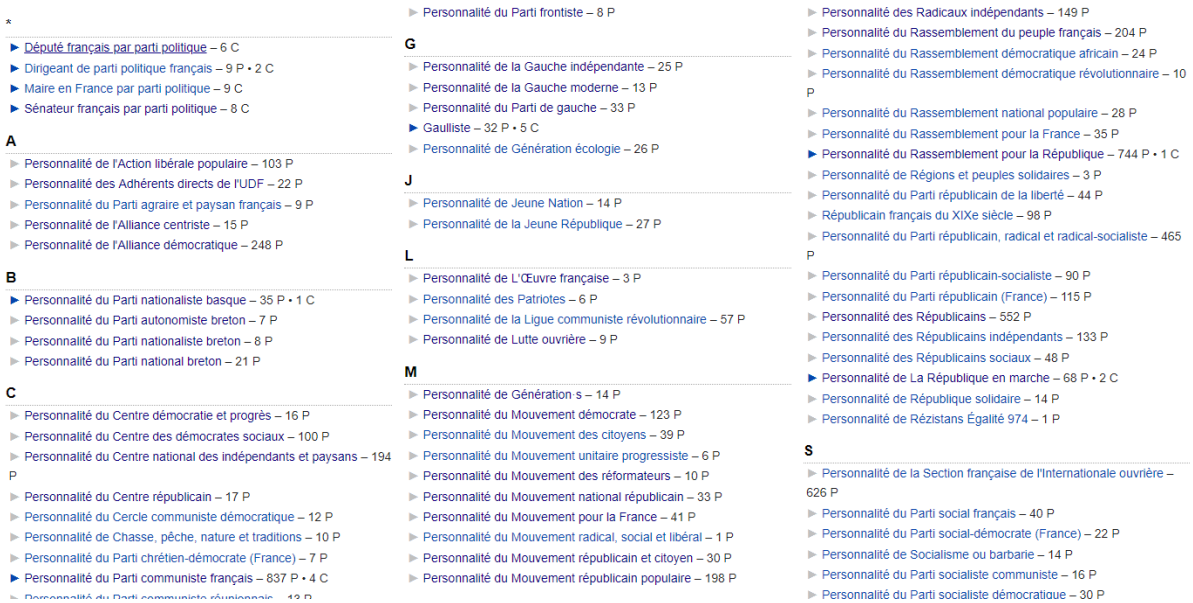
Afin de déceler les cas de répétition de la question, on pourrait faire appel à l’historique et vérifier si la liste des clés + attributs correspond à celle des clés + attributs de la question courante.

Partie Connaissance

Corpus de données

Le corpus de données est extrait à partir du script python extractionWiki.py dans le dossier Extraction.

Nous avons choisi d'obtenir nos données depuis la catégorie "Personnalité politique française par parti" :



*
- Député français par parti politique – 6 C
- Dirigeant de parti politique français – 9 P • 2 C
- Maire en France par parti politique – 9 C
- Sénateur français par parti politique – 8 C

A
- Personnalité de l'Action libérale populaire – 103 P
- Personnalité des Adhérents directs de l'UDF – 22 P
- Personnalité du Parti agraire et paysan français – 9 P
- Personnalité de l'Alliance centriste – 15 P
- Personnalité de l'Alliance démocratique – 248 P

B
- Personnalité du Parti nationaliste basque – 35 P • 1 C
- Personnalité du Parti autonomiste breton – 7 P
- Personnalité du Parti nationaliste breton – 8 P
- Personnalité du Parti national breton – 21 P

C
- Personnalité du Centre démocratie et progrès – 16 P
- Personnalité du Centre des démocrates sociaux – 100 P
- Personnalité du Centre national des indépendants et paysans – 194 P
- Personnalité du Centre républicain – 17 P
- Personnalité du Cercle communiste démocratique – 12 P
- Personnalité de Chasse, pêche, nature et traditions – 10 P
- Personnalité du Parti chrétien-démocrate (France) – 7 P
- Personnalité du Parti communiste français – 837 P • 4 C
- Personnalité du Parti communiste réunionnais – 13 P
- Personnalité du Parti frontiste – 8 P

G
- Personnalité de la Gauche indépendante – 25 P
- Personnalité de la Gauche moderne – 13 P
- Personnalité du Parti de gauche – 33 P
- Gaulliste – 32 P • 5 C
- Personnalité de Génération écologie – 26 P

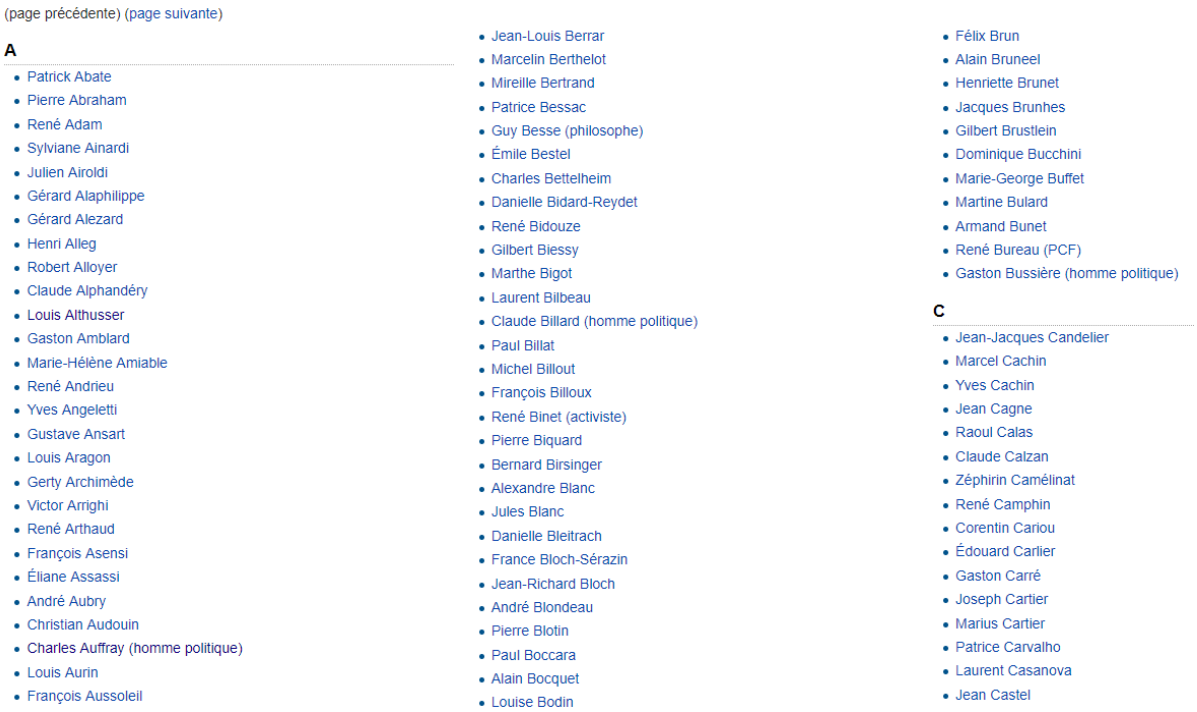
J
- Personnalité de Jeune Nation – 14 P
- Personnalité de la Jeune République – 27 P

L
- Personnalité de L'Œuvre française – 3 P
- Personnalité des Patriotes – 6 P
- Personnalité de la Ligue communiste révolutionnaire – 57 P
- Personnalité de Lutte ouvrière – 9 P

M
- Personnalité de Génération·s – 14 P
- Personnalité du Mouvement démocrate – 123 P
- Personnalité du Mouvement des citoyens – 39 P
- Personnalité du Mouvement unitaire progressiste – 6 P
- Personnalité du Mouvement des réformateurs – 10 P
- Personnalité du Mouvement national républicain – 33 P
- Personnalité du Mouvement pour la France – 41 P
- Personnalité du Mouvement radical, social et libéral – 1 P
- Personnalité du Mouvement républicain et citoyen – 30 P
- Personnalité du Mouvement républicain populaire – 198 P
- Personnalité des Radicaux indépendants – 149 P
- Personnalité du Rassemblement du peuple français – 204 P
- Personnalité du Rassemblement démocratique africain – 24 P
- Personnalité du Rassemblement démocratique révolutionnaire – 10 P
- Personnalité du Rassemblement national populaire – 28 P
- Personnalité du Rassemblement pour la France – 35 P
- Personnalité du Rassemblement pour la République – 744 P • 1 C
- Personnalité de Régions et peuples solidaires – 3 P
- Personnalité du Parti républicain de la liberté – 44 P
- Républicain français du XIXe siècle – 98 P
- Personnalité du Parti républicain, radical et radical-socialiste – 465 P
- Personnalité du Parti républicain-socialiste – 90 P
- Personnalité du Parti républicain (France) – 115 P
- Personnalité des Républicains – 552 P
- Personnalité des Républicains indépendants – 133 P
- Personnalité des Républicains sociaux – 48 P
- Personnalité de La République en marche – 68 P • 2 C
- Personnalité de République solidaire – 14 P
- Personnalité de Rézistans Égalité 974 – 1 P

S
- Personnalité de la Section française de l'Internationale ouvrière – 626 P
- Personnalité du Parti social français – 40 P
- Personnalité du Parti social-démocrate (France) – 22 P
- Personnalité de Socialisme ou barbarie – 14 P
- Personnalité du Parti socialiste communiste – 16 P
- Personnalité du Parti socialiste démocratique – 30 P

Ici, le script entre dans chacun des partis politique :



(page précédente) (page suivante)

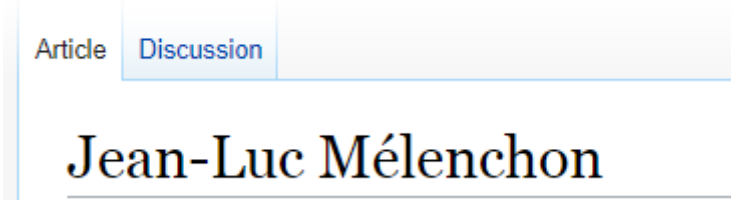
A
- Patrick Abate
- Pierre Abraham
- René Adam
- Sylviane Ainardi
- Julien Airoldi
- Gérard Alaphilippe
- Gérard Alezard
- Henri Alleg
- Robert Alloyer
- Claude Alphandéry
- Louis Althusser
- Gaston Amblard
- Marie-Hélène Amiable
- René Andrieu
- Yves Angeletti
- Gustave Ansart
- Louis Aragon
- Gerty Archimède
- Victor Arrighi
- René Arthaud
- François Asensi
- Éliane Assassi
- André Aubry
- Christian Audouin
- Charles Auffray (homme politique)
- Louis Aurin
- François Aussoleil
- Jean-Louis Berrar
- Marcelin Berthelot
- Mireille Bertrand
- Patrice Bessac
- Guy Besse (philosophe)
- Émile Bestel
- Charles Bettelheim
- Danielle Bidard-Reydet
- René Bidouze
- Gilbert Biessy
- Marthe Bigot
- Laurent Bilbeau
- Claude Billard (homme politique)
- Paul Billat
- Michel Billout
- François Billoux
- René Binet (activiste)
- Pierre Biquard
- Bernard Birsinger
- Alexandre Blanc
- Jules Blanc
- Danielle Bleitrach
- France Bloch-Sérazin
- Jean-Richard Bloch
- André Blondeau
- Pierre Blotin
- Paul Boccara
- Alain Bocquet
- Louise Bodin
- Félix Brun
- Alain Bruneel
- Henriette Brunet
- Jacques Brunhes
- Gilbert Brustlein
- Dominique Bucchini
- Marie-George Buffet
- Martine Bulard
- Armand Bunet
- René Bureau (PCF)
- Gaston Bussière (homme politique)

C
- Jean-Jacques Candelier
- Marcel Cachin
- Yves Cachin
- Jean Cagne
- Raoul Calas
- Claude Calzan
- Zéphirin Camélinat
- René Camphin
- Corentin Cariou
- Édouard Carlier
- Gaston Carré
- Joseph Cartier
- Marius Cartier
- Patrice Carvalho
- Laurent Casanova
- Jean Castel

Le script n'a plus qu'à s'introduire dans chacune des pages qui correspondent à des personnalités politiques.

D’une telle page, on va extraire nos différentes informations selon plusieurs méthodes.

Déjà, le nom et le prénom sont extraits du titre de la page Wikipédia :



Ensuite, les attributs non-formatés sont extraits depuis le texte. Par exemple :

Jean-Luc Mélenchon, né le 19 août 1951 à Tanger (Maroc), est un homme politique français.

Membre du Parti socialiste (PS) à partir de 1976, il est successivement élu conseiller municipal de Massy en 1983, conseiller général de l'Essonne en 1985 et sénateur en 1986. Il est également ministre délégué à l'Enseignement professionnel de 2000 à 2002, dans le gouvernement Lionel Jospin.

Il fait partie de l'aile gauche du PS jusqu'au congrès de Reims de 2008, à l'issue duquel il quitte le parti pour fonder le Parti de gauche (PG), dont il devient d'abord président du bureau national, puis coprésident, fonction qu'il conserve jusqu'en 2014.

Sous les couleurs du Front de gauche, coalition qui réunit notamment le PG et le Parti communiste (PCF), il est élu député européen en 2009 et réélu en 2014. Il est candidat de cette coalition à l'élection présidentielle de 2012, à l'issue de laquelle il arrive en quatrième position au premier tour, avec 11,10 % des voix. En 2016, il fonde le mouvement La France insoumise (FI) et se présente sous cette étiquette à l'élection présidentielle de l'année suivante, où il termine à nouveau en quatrième position, avec 19,58 % des suffrages exprimés.

Il est ensuite élu député dans la quatrième circonscription des Bouches-du-Rhône et préside le nouveau groupe La France insoumise à l'Assemblée nationale, au sein duquel il s'oppose à la politique du président Emmanuel Macron.

Ici on récupère notamment le lieu de naissance, la date de naissance et la particule associée à la personnalité politique.

On récupère les parties intéressantes, par exemple “Famille”, qui nous intéresse pour obtenir des informations sur la famille :

Biographie [modifier | modifier le code]

Famille [modifier | modifier le code]

Jean-Luc Mélenchon naît le 19 août 1951 à Tanger, au Maroc, alors zone internationale, où ses parents travaillent[1].

Il est le fils cadet de Georges Mélenchon, receveur des Postes, télégraphes et téléphones (PTT), et de Jeanine Bayona, institutrice, tous deux Français d'Algérie (« pieds-noirs »)[2, 3].

Son grand-père paternel, Antonio Melenchón est un Espagnol de la région de Murcie. Au début des années 1900, il s'installe à Oran, en Algérie française, et y épouse Aimée Canicio, elle aussi d'origine espagnole. Du côté maternel, son grand-père, François Bayonna, est né en 1889 près de Valence, en Espagne, et est marié à Jeanne Emmanuelle Caserta, une Italienne originaire de Sicile[2].

Jean-Luc Mélenchon reçoit une éducation catholique de par sa mère : il est notamment enfant de chœur et sert la messe en latin[3]. Sa mère s'éloigne ensuite de la religion[3]. Sans revendiquer sa foi, il déclare être de « culture catholique »[4].

En 1962, à la suite du divorce de ses parents deux ans plus tôt[5], il quitte le Maroc pour la France : il s'installe à Elbeuf, puis à Yvetot, dans le pays de Caux[6], puis dans le Jura, où sa mère est mutée[1, 7, 8].

Marié avec Bernadette Abriel durant son séjour à Besançon (puis divorcé), il a une fille, Maryline Camille[9], née en 1974[8], adjointe au maire du 4e arrondissement de Lyon de mars 2008 à février 2009[10] et compagne de Gabriel Amard, secrétaire national du Parti de gauche et ancien maire de Viry-Châtillon[11].

De même pour la formation.

Pour les données formalisées, la démarche est différente. Sur Wikipédia, beaucoup d'Hommes politiques se voient attribuer un cartouche qui résume les différents éléments de leur vie. Notamment, on retrouve les différents métiers liés à la politique exercés ainsi que les partis politiques :

Fonctions	
Président du groupe FI à l'Assemblée nationale	
En fonction depuis le 27 juin 2017 (8 mois et 5 jours)	
Législature	XV^e (Cinquième République)
Prédécesseur	*Création du groupe*
Député français	
En fonction depuis le 21 juin 2017 (8 mois et 11 jours)	
Élection	18 juin 2017
Circonscription	4^e des Bouches-du-Rhône
Législature	XV^e (Cinquième République)
Groupe politique	FI
Prédécesseur	Patrick Mennucci
Député européen	
14 juillet 2009 – 18 juin 2017 (7 ans, 11 mois et 4 jours)	
Élection	7 juin 2009
Réélection	25 mai 2014
Circonscription	Sud-Ouest
Législature	7^e et 8^e
Groupe politique	GUE/NGL
Successeur	Marie-Pierre Vieu
Président puis coprésident du bureau national du Parti de gauche	
1^er février 2009 – 22 août 2014 (5 ans, 6 mois et 21 jours)	
Avec	Martine Billard
Prédécesseur	*Fonction créée*

Parti politique	OCI (1972-1976) PS (1976-2008) PG (2008-2009) FG (2009-2016) FI (depuis 2016)

Ici la principale difficulté vient du fait que la structure HTML du cartouche est assez basique et ne permet pas une navigation structurée.

Extraction des données

L'extraction des données se fait depuis le fichier make_db.lua.

Données formatées

Les données formatées sont extraites avec des mots-clés de sorte à ce qu'il soit facile avec des patterns de retrouver les éléments.

Les données ont beau être formatées sur le papier, dans les faits la tâche est plus difficile. Typiquement pour les partis politiques, la syntaxe est très capricieuse.

nom (acronyme) (dates)

acronyme (nom) (dates)

nom (dates)

nom

etc...

Finalement les "patterns" sont nombreux et extraire ces données ne se révèle pas aussi simple que pour des données purement formatées.

Données non-formatées

Les données non-formatées sont reconnues à l'aide des différents outils à notre disposition par dark et lua.

Nous nous aidons d'un modèle statistique français afin notamment de détecter les noms communs et les noms propres.

Date de naissance et lieu de naissance :

La plupart des pages Wikipédia commencent de la même façon :

"[Homme politique] est né(e) le [date] à [lieu]". Il est donc aisé de les retirer afin de les ajouter à la base de données.

La particule est obtenue de manière extrêmement simple. Par défaut elle vaut "Il", car il existe plus d'hommes politiques que de femmes politiques.

Complétons le morceau de phrase analysé pour la date et le lieu de naissance. La plupart des pages Wikipédia commencent ainsi :

"[Homme politique] est né(e) le [date] à [lieu], est un(e) femme/homme politique".

Ainsi nous avons simplement à détecter c'"est une femme politique" afin de modifier la particule en "Elle".

Pour la formation, nous ne gérons que les cas du Baccalauréat, de la licence et du master. Les autres formations intéressantes ne sont malheureusement pas traitées car difficiles à extraire. Nous reviendrons sur le cas de la formation dans la partie couverture du rapport.

De cette formation nous essayons d'extraire le lieu, le sujet ainsi que la date d'obtention.

La famille quant à elle est extraite selon de nombreux patterns qu'il serait difficile d'expliciter dans le rapport. Le principal souci vient du fait que les mots-clés sont réciproques. Par exemple "père" peut très bien désigner le père de l'Homme politique, mais aussi le fait qu'il soit "père" d'un enfant.

Professions

Sur Wikipédia, les professions d'un Homme politique sont parfois très détaillées, avec par exemple le prédécesseur de la fonction, le successeur, le lieu, etc…

Cependant, la plupart de ces sous-attributs ne sont retrouvés qu'une ou deux fois dans tout le corpus.

Nous avons donc décidé de nous focaliser sur les plus communs : le nom de la profession, la date de début et la date de fin.

Dans l'extraction des données, nous avons décidé de stocker de manière dynamique tous les attributs d'une profession, bien que ces données ne soient pas encore utilisées par le système de dialogue.

Député français	
En fonction depuis le 21 juin 2017 (8 mois et 11 jours)	
Élection	18 juin 2017
Circonscription	4e des Bouches-du-Rhône
Législature	XVe (Cinquième République)
Groupe politique	FI
Prédécesseur	Patrick Mennucci

Ici par exemple, la table "profession" de "jean-luc_melenchon" ressemble à cela :

profession = {

intitule = "Député français",

date_début = "21 juin 2017",

date_fin = "",

election = "18 juin 2017",

circonscription = "4e des Bouches-du-Rhône",

legislature = "XVe (Cinquième République)",

Groupe_politique = "FI",

predecesseur = “Patrick Mennucci”,

}

On peut imaginer ces données exploitées avec un système de mot-clé ou autre. Malheureusement cette fonctionnalité n’est pas implantée coté Système de dialogue.

Couverture

Couverture totale

Afin d’évaluer la couverture totale, dans make_db.lua se trouve une fonction qui permet de calculer la couverture totale de chaque attribut de la base de données.

En voici le résultat commenté en italique :

Nombre de personnalités politiques totale : 7467

Couverture de 100% logique étant donné que la clé est générée à partir du nom et du prénom.

name : 7467 couverture : 100.0%

Couverture de 100% logique étant donné que la clé est générée à partir du nom et du prénom.

firstname : 7467 couverture : 100.0%

birth : 5685 couverture : 76.134993973483%

birthplace : 5939 couverture : 79.536627829115%

Par défaut la particule est à “Il”. C’est seulement si l’élément “femme politique” est présent dans le premier paragraphe Wikipédia que la particule est modifiée. Il faut donc ici regarder le taux d’erreur et non la couverture totale.

particule : 7467 couverture : 100.0%

famille : 813 couverture : 10.887906789875%

Nombre total de membres de famille : 1513

Nom : 642 couverture : 42.432253800397%

Prénom : 1513 couverture : 100.0%

Statut : 1513 couverture : 100.0%

Métier : 193 couverture : 12.756113681428%

parti : 4363 couverture : 58.430427213071%

Nombre total de partis : 7450

Nom : 7450 couverture : 100.0%

Acronyme : 4659 couverture : 62.536912751678%

Date adhésion : 497 couverture : 6.6711409395973%

Date départ : 497 couverture : 6.6711409395973%

profession : 5770 couverture : 77.273336011785%

Nombre total de professions : 12760%

Nom : 12760 couverture : 100.0%

Date adhésion : 8263 couverture : 64.757053291536%

Date départ : 6016 couverture : 47.147335423197%

formation : 420 couverture : 5.6247488951386%

Nombre total de formations : 455

Nom : 455 couverture : 100.0%

Sujet : 179 couverture : 39.340659340659%

Lieu : 48 couverture : 10.549450549451%

Date obtention : 105 couverture : 23.076923076923%

Couverture effective maximale

Etant donné le nombre assez élevé de politiciens dans la base de données, nous avons préféré nous focaliser sur un échantillon. Par manque de temps, cet échantillon est extrêmement restreint, étant constitué de 40 politiciens seulement.

Pour chaque politicien, on regarde si l'attribut est trouvable dans le corpus.

Nom : 40/40 = 100%

Prénom : 40/40 = 100%

Date de naissance : 33/40 = 82.5%

Lieu de naissance : 34/40 = 85%

Particule : 40/40 = 100%

Famille : 8/40 = 20%

Profession : 32/40 = 80%

Parti : 26/40 = 65%

Formation : 7/40 = 17.5%

Couverture

Nom : 40/40 = 100%

Prénom : 40/40 = 100%

Date de naissance : 33/33 = 100%

Lieu de naissance : 33/34 = 97.06%

Particule : 40/40 = 100%

Famille : 4/8 = 50%

Profession : 32/32 = 100%

Parti : 24/26 = 92.31%

Formation : 3/7 = 42.86%

Analyse de la couverture

Comme prévu, la couverture est beaucoup plus faible pour les données non-formatées que pour les données formatées.

Pour la famille, cela s'explique notamment par la difficulté à capter les informations. Sur Wikipédia, les façons de parler d'un membre de la famille sont extrêmement nombreuses et être exhaustif s'avère difficile.

Pour la formation, nous ne couvrons pour le moment que le baccalauréat, la licence et le master. Les autres diplômes et formations ne sont donc pas traités.

Afin d'améliorer la couverture de la formation, nous devrions créer un lexique des formations les plus connues afin de ne pas nous limiter à ces formations. Typiquement, une grande partie des politiciens font Science Po ou l'ENA.

Pour les partis politiques, nous ne traitons malheureusement que les données formatées. Par conséquent, lorsque ces dernières sont manquantes, il arrive qu'elles soient tout de même disponibles dans les données non-formatées, bien qu'à première vue ce cas soit assez rare.

Taux d'erreur

Pour certains attributs, le taux d'erreur était pertinent à évaluer.

Pour le lieu de naissance sur les 40 politiciens seul 1 indiquait un lieu totalement incorrect. Il arrive aussi que le lieu soit seulement approximatif, par exemple "Aix-les-Bains" devenant "Aix-". Cela est dû à l'utilisation du modèle statistique pour les noms propres. Nous devrions modifier le pattern de reconnaissance afin de prendre en compte les tirets et apostrophes par exemple.

Pour la famille, pas d'erreur sur les 40 politiciens, mais nous savons qu'il arrive que des erreurs soient introduites. Typiquement, la femme d'un homme politique peut être confondue avec la femme de son père.

Pour la particule, nous n'avons malheureusement pas eu le temps de tester sur un échantillon assez important pour détecter des erreurs. À priori, la très large majorité des femmes sont traitées avec la particule "Elle".

Améliorations à venir

Prise en compte des données non formatées pour les partis politique

Permet d'atteindre les 100% de couverture éventuellement

Permet de renforcer la base de données en indiquant par exemple si l'homme politique est le créateur du parti, un membre, etc…

Ajout d'autres formations pour l'extraction des données

Renforcement des patterns de famille

Prise en compte dans le système de dialogue des attributs dynamiques de profession.

Annexes : exemples de choses possibles avec le Système de Dialogue

“$help"

--- Cas Simple & Normal ---

"LIEU DE NAISSANCE ?","sep",

"Parlons de Laguiller s'il te plait.",

"quelle est sa date de naissance ?","sep",

"Laguiller et toi ?", "sep"

"de quel bord politique sont melenchon et macron ?",

"Quelle est la formation de Fillon ? et ses professions ?", "sep",

--- Cas Spécial ---

"qui sont les createurs d'ugoBot ?",

"qui sont tes createurs ?",

"et les miens ?", "sep",

"qui sont les créateurs de Macron et Mélenchon ?",

"Quelle est la réponse à la grande question sur la vie, l'univers et tout le reste ?",

"affiche moi l'historique", "sep",

--- Cas Complexe ---

"Lieu de naissance de Mélenchon et qui sont tes créateurs ?","sep",

"la formation et les partis de Melenchon ainsi que lieu de naissance de Macron ?",

"quels sont les partis auquels Melenchon et Macron ont été membre ?",

"quand Macron et melenchon ont-ils eu leur Baccalauréat ?",

--- Cas Limites du systèle de dialogue ---

"quelle est la date de naissance de Dominique ?",

"quand glotin a-t-il eu sa Licence ?",

-- exit

"au revoir et merci de votre attention ! :) ",

Vous trouverez ces exemples dans le fichier **bot.lua** dans la fonction **test_fonctionnel()**

Table des illustrations