NARCISO VIEGAS Mélanie
LEVEQUE Korantin
MICHARD Clément

Rapport de projet

PROJET POS-TAGGING

Table des matières

Introduction

Nous avons l'habitude de parler du POS-Tagging comme d'une tâche « terminée », c'est-à-dire d'une tâche dans laquelle les résultats optimaux ont déjà été obtenus. En effet, il se trouve que nous atteignons des scores de précision avoisinant les 95% : ce seuil indique qu'autant d'erreurs sont faites par une analyse automatique que par une analyse humaine. Nous pouvons cependant souligner quelques implications théoriques à ce problème, que nous soulèverons ici en tant que pistes de réflexion afin d'indiquer comment elles ont contribué à notre interprétation et à notre compréhension du problème, ainsi que donner un cadre pour quelques expériences complémentaires. Nous commencerons par un bref résumé de notre approche du problème, avant de détailler le protocole expérimental, puis nous finirons en résumant et explicitant nos tentatives d'expériences supplémentaires.

Analyse du problème

Ce rapport s'ouvre par une analyse et une réflexion autour de la tâche du POS Tagging en général, et de l'emploi du perceptron à cet usage en particulier. Il nous a semblé important d'intégrer ces expériences au reste de nos études, pour mieux contextualiser l'usage des probabilités pour le traitement automatique des langues. En effet, plutôt que d'appliquer simplement un algorithme et d'en tirer des résultats, nous avons souhaité le manipuler et apprendre à nous en servir, afin d'en faire un outil dans lequel nous pourrons nous fier. Cette démarche enfin

nous donne l'occasion de retracer un chemin, déjà parcouru, d'où l'on peut découvrir une perspective de la langue tout à fait différente de celle dont nous avons l'habitude. Deux questions nous ont semblé propres à l'approfondissement : au fond, qu'est-ce que l'étiquetage en parties du discours ? Et dans quelle mesure un tel étiquetage influe-t-il sur l'analyse qu'on fait des faits de langues ? Nous les aborderons d'abord à travers quelques notions historiques sur les parties du discours, puis sur la tournure spécifique qu'a prise cette pratique avec l'arrivée de l'ère informatique, pour finir par mentionner quelques conséquences de l'emploi d'un outil tel que l'informatique.

Quelques notions sur le POS-Tagging

Le POS Tagging, ou Part-of-Speech Tagging, est l'automatisation d'un procédé historiquement ancien. La philosophie aristotélicienne tentait déjà de distinguer des parties du discours - alors « catégories », qui comprenaient des éléments qui nous paraissent aujourd'hui disparates : ainsi verbe, nom, mais aussi syllabe étaient des catégories. Cependant l'aspect plus habituel et contemporain, qui associe un mot en tant qu'unité de phrase, à une partie du discours, a mûri dans les mains des grammairiens latins (Varon, Donatus, Denys le Thrace...). Cet esprit classificateur et normalisateur s'est maintenu comme une tradition fermement ancrée dans la conscience sociolinguistique des nations Européennes.

Il aura fallu un contact prolongé de la civilisation occidentale avec des phénomènes linguistiques qui ne répondaient pas aux critères usuels pour que naisse et se délimite clairement le besoin de sortir la grammaire d'un mode normatif, et d’étudier à quel point elle était valide d'un point de vue descriptif. Les premiers temps de la linguistique étaient ceux de la grammaire comparée, de Jones et des paradigmes si proches entre latin et sanskrit – mais une révolution s'est amorcée avec les premiers linguistes américains, qui voulaient étendre le modèle grammatical aux langues amérindiennes. La classification – autant génétique que typologique - de ces langues a longtemps été purement impossible. De la même façon, les premières descriptions des langues exotiques font des caricatures des langues européennes, dont il manquerait quelque chose : par exemple, le chinois

qui procède d'une typologie isolante, a longtemps été considéré comme une langue de singes en ce qu'elle apposait simplement des mots aux mots, sans accord aucun.

La classification typologique des langues amérindiennes s'est donc confrontée à la relative opacité du système grammatical indo-européen : les paradigmes qui pouvaient s'y présenter n'étaient pas compatibles avec ceux des langues connues par la linguistique de cette époque. On peut par exemple rappeler la langue Hopi chère à Whorf, où un certain nombre de notions étaient absentes ou séparées dans les paradigmes de flexions des différentes catégories.

La méthode de Bloomfield aura permis de rationaliser l'étude des langues amérindiennes : dans le courant du structuralisme, il définissait la syntaxe comme les séquences possibles de catégories morphosyntaxiques. Cette méthode a été la première à supposer un travail sur de grands corpus pour l'étiquetage des parties du discours ; et c'est aussi celle qui a pu donner des résultats probants, et définir des paradigmes solides pour ces langues qui semblaient jusqu'alors fort éloignées de cette entreprise de classification.

Voyons à présent ce que ces considérations générales ont apporté à notre étude particulière.

De l'intérêt du perceptron pour le POS-Tagging

Si les premières annotations et les premiers traitements de corpus furent manuels, l'ère de l'informatisation a permis de rendre ce processus réellement accessible et applicable. Plusieurs algorithmes qui visaient à bien délimiter les tâches en jeu, à les formaliser et à les optimiser, ainsi que l'augmentation croissante des ressources et des performances de l'ordinateur ont même permis aux alternatives, dans le traitement de corpus conséquents, de s'épanouir et de fleurir. Nous pouvons par exemple citer les algorithmes k-means, les réseaux neuronaux, dont le perceptron est un exemple minimaliste, ou encore les variations sur le modèle de Markov caché : les possibilités sont vastes.

Nous nous sommes arrêtés sur le choix d'un perceptron suite à l'analyse du problème du point de vue linguistique. En effet, la condition première de la classification par perceptron est que les classes soient linéairement séparables. Si l'on suppose que les données linguistiques sont placées dans un hyper-espace, alors le perceptron va définir un jeu d'hyperplans sécants qui chacun permettront de distinguer une classe des autres. Si l'on transpose cette condition d'ordre mathématique en des termes plus proche du vocabulaire de la linguistique, Nous obtenons un perceptron qui peut classifier des mots en parties du discours pour peu que les mots puissent être distinctement séparés selon leur contexte d'occurrence. En effet les « features » du perceptron étant purement aveugles – dans le sens où elles ne peuvent être influencées par d'autres abstractions que celles que le linguiste aura supposé pertinentes – la classification qu'il met en œuvre est d'ordre purement combinatoire : les mêmes données, les mêmes tests, les mêmes calculs sont répétés et auront toujours la même conséquence, le même impact, le même résultat. Nous retrouvons par là même la linguistique distributive de Bloomfield ; en ce qu'il s'agit de donner des traits pour classifier, et que cette classification s'établit ensuite de manière mécanique suivant ces traits.

Considérations sur les limites informatiques

Cependant cette condition algorithmique de classes linéairement séparables est, à bien y penser, une contrainte forte. Si on s'intéresse par exemple au nahuatl, qui relève d'une typologie presque polysynthétique, la question d'une classification morphosyntaxique devient beaucoup plus malaisée. La plupart des noms peuvent très naturellement prendre la fonction de prédicat (ce que Michel Launay nommait l'omniprédicativité) ; et en tant que prédicats ils sont susceptibles, tous comme les verbes, de prendre des compléments ; de plus les paradigmes de conjugaisons personnelles s'appliquent autant à un verbe qu'à un nom. On n'y distingue, à vrai dire, un nom d'un verbe que par l'affixation de mar-

queurs TAM généraux chez les premiers, et la distinction entre un mode possédé et un mode absolu pour les seconds.

Un tel exemple montre bien que la séparation linéaire par « features » des mots peut prendre un caractère hypothétique. C'est seulement la multiplication des « features » qui permettra peut-être de séparer ces ensembles morphosyntaxiques. Cette multiplication a un prix : celui de la complexité – autant la complexité algorithmique que la complexité théorique sur le plan linguistique. Dans ce cadre-ci, nous soulignerons plus particulièrement que les calculs que suppose une telle approche nécessitent des ressources importantes. Aussi nous sommes-nous heurtés à des problèmes de complexité algorithmique alors que nous voulions prolonger nos expériences. Il aurait été nécessaire pour les mener à bien, d'une part, de réviser l'algorithme afin de supposer une complexité en espace moindre, d'autre part, d'utiliser des outils moins coûteux que Python.

Nous reviendrons plus particulièrement sur ces points en partie III de ce rapport.

Résumé des expériences

Une fois nos préliminaires théoriques posés, nous nous sommes attelés aux expériences proprement dites. Ayant choisi d'implémenter le tagger à l'aide d'un perceptron, il nous fallait donc arrêter celui des « features » exactes que nous allions employer ainsi que du nombre de tours pour entraîner le perceptron. Nous détaillerons le choix des « features » dans une première partie. De plus, le but de cette expérience étant de comparer les différences induites par des tagsets plus ou moins fins, nous avons décidé d'instancier trois perceptrons : deux avec des tags universaux, l'un français, l'autre allemand, et le dernier avec des étiquettes fines sur le corpus tiger en allemand. Cette décision sera explicitée dans un second temps. Enfin nous donnerons les résultats et les interprétations de ceux-ci dans une troisième sous-partie.

« Features »

Pour cette étude nous avons choisi le perceptron multi-classe moyenné. Nous sommes donc partis à la fois de celui étudié en cours ainsi que de celui fourni pour la réalisation de ce travail. Plus précisément, nous avons supposé que plusieurs facteurs étaient à même de jouer un rôle dans la classification linguistique. La première distinction était celle entre faits d'ordre morphologique et faits d'ordre syntaxique. Pour modéliser l'impact de la morphologie, sans pour autant vouloir présupposer une analyse morphologique, le plus simple étaient d'étudier le suffixe et le préfixe de chaque mot, puisque ceux-ci sont susceptibles de porter des marques de dérivation morphologique ; sachant de plus que les marques flexionnelles sont portées en fin de mot en français et en allemand, et définissent le plus souvent le paradigme auquel se rattache le mot, il nous a semblé préférable d'étudier plus en détail le contexte suffixé. Aussi avons-nous retenu l'emploi comme « features » le suffixe de trois caractères et le préfixe de deux caractères de chaque mot.

En regard des faits morphologiques, il faut tenir compte de la syntaxe – à proprement parler du contexte d'occurrence des tokens, puisqu'une analyse syntaxique présuppose un étiquetage morphosyntaxique. Ceci s'inscrivait de plus dans notre réflexion sur la méthode de Bloomfield, qui consistait à étudier les classes que l'on peut dresser selon le contexte d'occurrence des mots. Par conséquent nous avons conservé le fait d'utiliser comme « features » les deux mots précédents (chacun comptant pour une « feature ») et les deux mots suivants (comptant pareillement).

De plus, les catégories morphosyntaxiques sont déterminées l'une par l'emploi des autres : par exemple un Nom sera souvent précédé d'un Déterminant, inversement, un Déterminant sera suivi d'un Nom ; le sujet d'un Verbe étant en français presque exclusivement nominal, on s'attend à trouver, parmi les mots précédent celui-ci, un Nom. Ces faits sont rendus moins clairs par la possibilité d'entremêler le groupe nominal, par exemple, d'adjectifs – ou le groupe verbal, pareillement, d'adverbes. Il y a là en vérité deux faits : d'une part, il existe

des séquences licites de catégories du discours, et d'autre part, certaines catégories impliquent fortement l'existence, dans un contexte proche, de catégories d'une autre classe. Pour modéliser le premier fait, Nous choisissons de conserver la séquence des deux tags précédents. Pour modéliser le second, puisque l'attribution de tags se fait de manière séquentielle, nous ne pouvons nous baser que sur le contexte précédent ; par conséquent les deux tags qui précèdent le mot actuel font chacun l'objet d'une « feature ».

Le mot lui-même, ainsi que sa forme normalisée servent aussi traditionnellement de « features » : catégoriser le mot lui-même comme une « feature » permet de pondérer conséquemment le vecteur associé au mot afin que le même mot soit a priori classé toujours de la même façon. Cependant, pour permettre de contrebalancer cette pondération, et laisser une possibilité à des homographes d'être classés différemment, la forme normalisée du mot (forme canonique du lexème) est aussi employée comme « feature ».

Enfin, afin de satisfaire à des exigences mathématiques, le biais a aussi été considéré comme une « feature ».

Soit par conséquent la liste de « features » suivante :

- un biais (constante)
- le mot
- la forme
- un suffixe (trois caractères)
- un préfixe (deux caractères)
- le premier mot avant
- le deuxième mot avant
- le premier mot après
- le deuxième mot après

- le premier tag précédent
- le deuxième tag précédent
- les deux derniers tags

Tag-sets

L'un des points importants pour ces expériences était l'utilisation des ressources mises au point par Petrov. L'article de Petrov, Das, et McDonald introduit un jeu d'étiquettes universelles, qui, couplées avec des mappings vers les jeux d'étiquettes de corpus plus spécialisés, permettent notamment de comparer entre elles deux analyses en catégories morphosyntaxiques. De manière générale, Petrov et ses collègues souhaitent établir un tag-set qui soit pragmatiquement assez proche des besoins des linguistes, et assez général pour permettre une utilisation de manière universelle.

Notre analyse de ce TP, ainsi que l'article de Petrov et al., laisse penser que la définition de telles catégories est évidemment sujet à débats. L'idée de Petrov était d'établir un mapping vers des classes assez générales pour un certain nombre de langues en corpus doté d'un tag-set précis. Il illustre cette démarche avec un exemple issu du PennTreeBank : les catégories VB, VBD, VBG, VBN, VBP, VBZ, et MD sont subsumées par la catégorie 'VERB' du tag-set universel. Si chacune des tags originaux du PennTreeBank peuvent être délimités par leur contexte, c'est-à-dire si chacune des classes est linéairement séparable par un hyperplan, alors on peut penser que les catégories regroupées dans le tag-set conserveront les mêmes propriétés (le contexte spécifique serait alors l'union des deux contextes). Aussi sommes-nous partis de l'opinion qu'un tag-set plus riche et un tag-set moins fin, devraient donner des résultats similaires.

Nous soulignons ici deux faits supplémentaires concernant les tag-sets : premièrement que le tiger corpus, sur lequel se sont basées certaines de nos expériences, est l'un des deux corpus (avec le Negra) qui ont servi à l'élaboration du tag-set universel. Ceci implique un parallélisme important entre le tag-set universel et celui, plus fin, du tiger corpus. Secondement, probablement à cause de

la disparité dans l'accessibilité des ressources, le tag-set universel est très fortement influencé par des corpus indo-européens. Les seules langues non indo-européennes qui y étaient représentées lors de la rédaction de l'article de Petrov et al. sont le chinois, le basque, le hongrois, le turc, le japonais et le coréen. À l'exception du premier, toutes ces langues procèdent de modèles morphosyntaxiques très similaires. On notera l'absence d'un certain nombre de typologies et de familles de langues – pas de langues polysynthétiques, pas de créoles ; on peut résumer cet ensemble de langue en disant que ces langues possèdent toutes une grammaire « forte ».

Ces différentes remarques – autant les deux précédentes que celles issues de notre analyse préliminaire –, nous les avons résumées ainsi : la séparabilité du vocabulaire d'une langue en catégories du discours est très certainement une caractéristique typologique propre aux langues à grammaires fortes. Il faut donc justifier, d'une part, que cette séparation en catégorie est unique, d'autre part, que cette séparabilité est universelle et non pas l'apanage d'un groupe de langues seulement.

Par conséquent nous avons choisi d'instancier trois perceptrons. Le premier s'entraînait sur un corpus français avec les étiquettes universelles de Petrov. À titre de comparaison quant à l'efficacité de ce tag-sets à travers les langues, nous avons instancié un second perceptron, classant à l'aide des mêmes étiquettes, sur un corpus allemand. Enfin, pour étudier l'efficacité spécifique d'un jeu plus fin contre un jeu plus universel d'étiquettes, notre troisième perceptron a été instancié à partir des données du corpus tiger. Nous avons souhaité étudier l'impact que ces différents classements pourraient avoir ; en nous aidant d'une part d'une matrice de confusion pour nos analyses, d'autre part en observant l'impact sur les performances lors d'une série d'opérations – ici nous avons enchaîné à la classification une analyse en dépendance.

Résultats

Nous disposions de trois groupes de corpus et d'une méthode de calcul de précision (un rapport simple entre tags prédits et tags annotés). De plus, pour

analyse plus fine des données, nous avions à notre disposition la matrice de confusion, implémentée en tant qu'objet autonome instancié depuis les données formatées par la fonction read_conll.

Nos expériences se divisent en deux parties, d'abord l'étiquetage morpho-syntaxique, le cœur de l'étude et dans un second temps, le parsing en dépendances, utilisant, les différents étiquetages appris sur les données gold.

Tout d'abord, nous avons fixé à dix tours, le nombre d'itérations du perceptron pour toutes les expériences.

Nous avons en premier lieu entraîné un premier perceptron sur le Tiger Corpus, soit un corpus à étiquetage fin. Cette expérience s'est soldée par un score de 97,6 % de réussite. Après cela, nous avons entraîné un deuxième perceptron mais cette fois-ci sur un corpus à étiquetage universel mais toujours sur de l'allemand. Nous avons obtenu 94,5 % de réussite.

En observant à la fois cette différence de 3,1 % et du temps qu’on mis les deux perceptron à apprendre les données, nous avons mis en place une matrice de confusion sur les deux résultats obtenus. L'interprétation de la matrice nous a permis de fusionner certaines catégories qui étaient plus souvent mal étiquetées que bien étiquetées. Après ces fusions catégorielles sur le corpus d’entraînement, nous avons entraîné à nouveau le perceptron, mais ces dernière modifications n'ont changé grand-chose car nous nous sommes simplement fixés sur les chiffres de la matrice, nous n'avons fusionner des catégories que s'il y avait un mauvais étiquetage.

En conclusion de cette première expérience, malgré un cout de temps plus important, il est tout de même bénéfique d'avoir un étiquetage fin, même si, fondamentalement, les différences ne sont pas énormes.

A partir des résultats du POSTagging, nous avons passé les données étiquetés dans un analyseur en dépendances. Par rapport aux chiffres sortis, nous pouvons dire qu'un étiquetage universel est amplement suffisant pour parseur des données. Nous obtenons à 0,2 % près les mêmes résultats que le POSTagging soit fin ou pas. Le tagueur va aussi plus vite sur des données à tags universels. En conclusion, autant rester sur une palette d'étiquettes raisonnables tout en conservant un temps de réaction et une qualité suffisante.

Expériences supplémentaires

Nous souhaitions par la suite étudier si les mêmes classes pouvaient émerger d'elles-mêmes des données – c'est-à-dire s'il était inhérent à la langue que les mots fussent séparés selon tel ensemble de catégories. Nous avons d'abord cherché à voir si les confusions dans le classement du perceptron pouvaient tendre à une répartition similaire au tag-set universel, et, à l'inverse, si n'importe quel ensemble de tags pouvait prétendre à partitionner les données de la langue.

Expérience 1 : « auto-structuration »

Nous avons envisagé tout d'abord de voir, à partir de classes singletons (ne se rapportant qu'à l'occurrence d'un seul token) et des regroupements suggérés par la matrice de confusion, quel tag-set nous pouvions obtenir. Par conséquent, nous avons retaggé le corpus d'entraînement du perceptron, en associant à chaque occurrence de mot une classe qui lui est unique. Nous nous doutions que la complexité algorithmique d'un tel procédé soit pour le moins catastrophique – le perceptron demande de pondérer le vecteur poids en tenant compte de chacune des classes possibles, or, dans notre cas il y en avait autant que d'occurrences de mots dans le corpus. Il est donc tout à fait attendu que nous n'avons pas en mesure de mener une telle expérience à bien. Nous avons cependant tenté de réaliser cette expérience tout d'abord sur le corpus entier, pour chaque occurrence ; ensuite, en réduisant la taille du corpus, puis en substituant aux classes singleton d'occurrence des classes regroupant les homographes. Le problème était ici d'ordre purement algorithmique ; il nous aurait fallu utiliser un algorithme visant à structurer des données encore informes. Nous pensons qu'une variation d'un algorithme k-means, qui procéderait par exemple par réunions d'ensembles de points dans l'espace vectoriel des vecteurs de mots plu-

tôt que par déplacement du centre du cluster aurait pu être un algorithme plus adapté pour une telle expérience.

Expérience 2 : « Random tag-set »

Notre deuxième série d'expériences personnelles a été de voir dans quelle mesure n'importe quel ensemble de tags pouvait partitionner le vocabulaire en classes. Pour ce faire, nous avons décidé de modifier le corpus d'entraînement afin d'attribuer des classes aléatoires aux mots du corpus ; nous souhaitions voir quelle précision atteindrait un perceptron, et quelles confusions il pourrait commettre. Cette expérience était partie d'une prise en compte la nécessité d'employer parcimonieusement les ressources informatiques dont nous disposions.

Dans une optique purement observationnelle, nous avons ré étiqueté de manière aléatoire le corpus avec des tags fantaisistes. Nous avons donc choisi huit classes « a, b, c, d, e, f, g et h » que nous avons démultipliées et mélangées. Puis de manière complètement aléatoire (enfin, plutôt une manière pseudo aléatoire que peut nous offrir le module Random de python), nous avons ré étiqueté le corpus du français taggé universel.

Cette expérience montre simplement que l'outil du perceptron est « idiot » ou du moins qu'il est purement mathématique. Avec cette redéfinition des tags, il nous sort 97,6 % de réussite.

L'échec paraît logique, du fait qu'on ait fixé des « features » en lien à la fois à la forme du mot, de son contexte mot à mot et tag à tag, le fait de faire un étiquetage complètement aléatoire, on perd nécessairement une foule d'information et un mot nom ambigu tel que « parlement » par exemple deviendra peut-être ambigu avec cette annotation.

Conclusion

Ces différents jeux d'expériences – d'une part, l'entraînement classique d'un perceptron sur des données du français et de l'allemand, et d'autre part, nos tentatives propres d'usage du perceptron pour structurer l'ensemble amorphe des occurrences non-taggées – nous indiquent très clairement que les langues à grammaires fortes ont la propriété de pouvoir séparer les mots selon des catégories du discours ne correspond pas seulement à une propriété du lexique. Les erreurs de classification produites par l'emploi d'un jeu de tags aléatoires mettent en évidence la dimension combinatoire forte qui existe dans ces langues : cette dimension syntaxique était l'intuition qui sous-tendait la démarche de Bloomfield, et c'est la même intuition que partagent les locuteurs de ces langues. Nous n'avons malheureusement pas les moyens d'étendre cette étude aux cas extrêmes et atypiques que constituent les langues polysynthétiques, les pidgins, les créoles. Aussi ne savons-nous pas trancher si la notion de « grammaire forte » est à proprement parler une caractéristique typologique de certaines langues, ou si nous avons là un réel universel du langage humain. Il serait intéressant de confronter, de plus, des classifications produites par d'autres algorithmes à celles issues de nos perceptrons : nous pourrions avancer plus sûrement que le tag-set propre à une langue donnée, ou le tag-set universel de Petrov reflètent en partie la même structuration du langage. De plus, comme nous l'avons suggéré plus haut, une variation sur un algorithme d'apprentissage par cluster pourrait répondre à notre question de si ce tag-set est contenu dans les données, ou s'il faut y voir une abstraction valide mais d'origine extérieure aux données. Ce projet nous aura permis d'aborder la linguistique selon deux perspectives opposées, et de voir où elles se rejoignent : la perspective d'une linguistique régulière, qui donne par exemple les lois combinatoires de la syntaxe, et la perspective d'une linguistique d'étude de corpus, telle que nous propose un algorithme de classification comme le perceptron. Ces deux perspectives confluent dans ce que les mêmes faits saillants sont mis en reliefs, même par des méthodes différentes : elles soulignent qu'il existe une série de conditions pour qu'une collection de

mots puisse être appelée langue, conditions que ne satisfait pas nécessairement n’importe quel ensemble.