
BICH-NGOC HOANG
PIERRE PRABLANC
YANG YANG

# Extraction d'Information dans des Textes

**Université Lyon 2**

**M2 Data-Mining 2018-2019**

Professeur référent :
Julien Ah-Pine (**julien.ah-pine@univ-lyon2.fr**)

# Table des matières

# Chapitre 1

# Introduction

## 1.1 Introduction Générale

Dans ce projet, nous nous proposons de traiter la tâche de reconnaissances d'entités nommées qui est une sous-tâche de l'extraction d'informations. La reconnaissance d'entités nommées consiste à détecter une entité textuelle (un mot, ou un groupe de mots) et à la classer dans une catégorie pouvant être des noms de personnes, de lieux, d'organisations, ou d'autres catégories plus spécifiques. Dans notre cas, nous nous intéressons à la reconnaissance de noms communs issus de la littérature scientifique médicale.

La quantité ainsi que la production de littérature dans les domaines scientifiques connaît une telle croissance qu'il est devenu humainement difficile d'analyser les articles pertinents noyés dans la masse de documents. La reconnaissance d'entité nommées permet donc d'aider à sélectionner les documents contenant des noms appartenant à une certaine entité (comme la catégorie des protéines par exemple). Les systèmes d'extractions de connaissance, qui servent à réunir et mettre à disposition de manière structurée l'information disponible dans la littérature, sont également alimentés par la reconnaissance d'entités nommées.

## 1.2 Problématique

TODO:Le problème considéré est une tâche de catégorisation pour laquelle il faut attribuer une classe à un mot. Nous disposons pour cela d'un corpus de données annotées dont une partie doit servir à alimenter le système de reconnaissance tandis que l'autre sert à l'évaluer.
Pour résoudre ce problème, on pourrait penser que la solution triviale consiste à comparer le mot à classifier à des dictionnaires de mots obtenus dans le jeu de données labellisées. Chaque dictionnaire correspondant à une classe, si le mot existe dans un dictionnaire alors il serait attribué à la classe correspondante. Une telle solution trouve malheureusement rapidement de nombreuses limitations. Comment classifier un mot qui n'existe dans aucun des dictionnaires (c'est à dire jamais observé dans le jeu d'entraînement) ? Par ailleurs, un même mot peut recouvrir plusieurs sens et donc appartenir à une classe différente selon le contexte. Il s'agit ici d'un problème d'ambiguïté. Le problème n'étant pas si simple, plusieurs méthodes ont été proposées.

## 1.3 Différentes Approches

### Méthodes Rules-Based

Historiquement, les premières méthodes étaient basées sur des règles ("Rules-Based") élaborées soit de manière automatique, soit écrites "à la main". Le principe se base sur des paires $(pattern, action)$ où

un $pattern$ correspond généralement à un une expression régulière. Lorsqu'un token (ou une séquence de tokens) correspond à un $pattern$, une $action$ associée est exécutée. Cette $action$ correspond à l'étiquetage des tokens (entité, début ou fin de l'entité par exemple). Bien que très efficaces, ces approches nécessitent, dans le cas des règles écrites à la main, un expert à la fois de la langue et du domaine, rendant la technique très spécifique et très coûteuse en temps. Pour le cas des règles obtenues de manière automatique, elles souffrent d'un manque de précision [Mladenić, 2017]. On trouve également d'autres inconvénients impactant par exemple la robustesse du système de reconnaissance. En effet, lorsque de nouvelles données nécessitent de nouvelles règles, il faut alors mettre à jour la table de règles.

**Méthodes Statistiques**

Les approches statistiques permettent de palier certains inconvénients des méthodes Rules-Based car les "règles" sont apprises sur les données. On distingue dans ces méthodes d'apprentissage statistique 3 types d'approches : non-supervisées, semi-supervisées et supervisées.

Machine-learning techniques Supervised-Learning : (non exhaustif) SVM, CRF, HMM, Naïve Bayesian, Neural Network, Decision Tree, Maximum Entropy Model Semi-Supervised-Learning : (données labélisées et non labélisées) Boot-strapping, Co-Training, Unsupervised-Learning : méthodes de clustering. Dans notre étude, nous ne nous concentrons que sur 3 méthodes d'apprentissage supervisé.

Développement de méthodes statistiques qui ont très rapidement dépassé les résultats des méthodes basées sur les règles.

## 1.4 Organisation du Rapport

TODO: On va faire de la reconnaissance d'entités nommées sur le Corpus GENIA. On va utiliser 3 méthodes (NB, HMM, CRF). Dans le chapitre 2, description d'une vue d'ensemble d'un système de reconnaissance d'entité nommée en détaillant ensuite chaque élément du système (pré-processing, extraction des descripteurs, méthodes de classifications utilisées) Dans le chapitre 3, on présente le corpus utilisé et les différentes configurations du système utilisé.

# Chapitre 2

# Reconnaissance d'Entité Nommées

## 2.1 Vue d'Ensemble

TODO:Faire un schéma à 2 colonnes (ou 2 lignes) :
Entrainement
Pre-processing -> Extraction Features -> Entrainement -> Cross-Validation
Test
Pre-processing -> Extraction Features -> Prédiction

## 2.2 Pré-Traitements

- Dans les pré-traitements, on part d'un corpus annoté, il y a donc une phase de prétraitement du corpus (généralement structuré dans un format balisé XML).
- extraction des entités nommées associées à leur label
- Les données possèdent parfois plusieurs labels, dans notre tâche de classification, on fait un choix sur le nombre de classes. On regroupe des classes entre elle (ontologie).
- Tout ça va être expliqué dans le chapitre "Expériences"

## 2.3 Extraction des Descripteurs

- Dans la tâche de reconnaissance d'entité nommée, on retrouve les descripteurs habituels en text mining (tokens, lemmes, pos) mais d'autres descripteurs moins courant comme les special verb trigger, word formation pattern
- Les descripteurs ont une importance capitale dans cette tâche de classifiation. Les résultats sont très dépendant du choix des descripteurs. c'est ce que nous verrons dans les expériences.

## 2.4 Méthodes de Classification

### 2.4.1 Bayésien Naïf

Les modèles bayésiens sont des modèles de classification qui se basent sur la règle de Bayes.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$ est appelée la probabilité a posteriori. $P(B|A)$ est la vraisemblance. $P(A)$ est la probabilité a priori et $P(B)$ est appelé l'évidence.

Appliqué à notre cas :
$P(c_i|X) = \frac{P(c_i)P(X|c_i)}{P(X)}$
Modèle bayésien cherche à maximiser la probabilité a posteriori
$\underset{x}{\operatorname{argmax}} P(c_i|X)$
Cela revient à maximiser le produit de la vraisemblance et de la probabilité a priori des classes :
$P(c_i|X) \propto P(c_i)P(X|c_i)$
$P(X|c_i) = P(x_1, x_2, ..., x_n|c_i)$
Hypothèse naïve entre les variables :
$P(X|c_i) = \prod_{j=1}^{n} P(x_j|c_i)$

$\underset{x}{\operatorname{argmax}} P(c_i|X) \propto P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$

Il faut donc calculer $P(c_i)$ et $P(x_j|c_i)$ pour tous les $j$. Pour le moment, aucune hypothèse sur la distribution utilisée modéliser la vraisemblance des données. Dans le cas catégoriel, on peut soit utiliser une distribution de Bernoulli multivariée soit une distribution multinomiale.

### 2.4.2 Hidden Markov Model

### 2.4.3 Conditional Random Field

# Chapitre 3

# Expériences

## 3.1 Corpus GENIA

TODO: Corpus GENIA : corpus développé pour le projet GENIA dont l'objectif est d'explorer les méthodes d'extraction d'information et de text mining spécifiques au domaine de la science médicale. Le corpus peut ainsi servir de référence pour la communauté scientifique dans les tâches citées ci-dessus. Le corpus GENIA est composé d'un sous-ensemble d'éléments d'articles de la base de données Medline spécifiques aux réactions biologiques impliquées dans les « transcriptions factors in human bloods cells ». Ainsi pour chaque article, seuls les titres et résumés ont été collectés à partir de requêtes sur l'interface web de PubMed.

Pré-traiter les documents du corpus xml de façon à enlever les méta-données et à représenter les textes sous forme numérique selon les besoins des classifieurs

## 3.2 Sélection des Catégories d'Entités Nommées

## 3.3 Méthodes de Classifications Testées

# Chapitre 4

# Discussions

# Conclusion

# Annexe A

# Glossaire

Main notations and usual acronyms included in this report are summarized thereafter.

| | |
|---|---|
| **A/S** | Analysis-Synthesis |
| **ASR** | Automatic Speech Recognition |
| **DTW** | Dynamic Time Warping |
| **EM** | Expectation Maximization algorithm |
| **F0** | Fundamental Frequency |
| **FT** | Fourier Transform |
| **GMM** | Gaussian Mixture Model |
| **HMM** | Hidden Markov Model |
| **HNM** | Harmonic plus Noise Model |
| **GV** | Global Variance |
| **LP** | Linear Prediction |
| **LPC** | Linear Prediction Coefficient |
| **LSF** | Line Spectral Coefficient |
| **LSP** | Line Spectral Pairs |
| **MFCC** | Mel Frequency Cepstral Coefficient |
| **MCD** | Mel-Cepstral Distortion |
| **NMF** | Non-negative Matrix Factorization |
| **PSOLA** | Pitch Synchronous Overlap Add |
| **PSOLA-NB** | Pitch Synchronous Overlap Add Narrow Band |
| **PSOLA-WB** | Pitch Synchronous Overlap Add Wide Band |
| **VC** | Voice Conversion |
| **STFT** | Short Time Fourier Transform |
| **STRAIGHT** | Speech representation and TRansformation using Adaptive Interpolation of weiGHTed spectrum |
| **TTS** | Text-To-Speech |
| **VQ** | Vector Quantization |

## Annexe B

# Algorithmes

# Bibliographie

[Mladenić, 2017] Mladenić, D. (2017). *Text Mining*, pages 1241–1242. Springer US, Boston, MA.