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**)

Contents

Chapter 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 peut donc 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.

[Greenwade, 1993]

1.2 Problématique

TODO:Problème : en se basant sur des données annotées, on essaie de retrouver des entités nommées dans de nouveaux documents.

Solution stupide : Faire une recherche des formes lémmatisées à partir d'un dictionnaire obtenue sur le jeu d'entrainement.

==>Impossible de retrouver de nouveaux mots
==>On risque de classer un mot figurant dans les mots annotés mais ayant un sens différent car dans un contexte différent.

Plusieurs méthodes ont été proposées:

Rules-based techniques Premières méthodes basées sur des règles faites à la main. Ces approches sont très efficaces. Cependant le problème est qu'elles nécessitent un expert à la fois de la langue et du domaine rendant la technique très spécifique et très coûteuse en temps. Règles comme « si 'X' est précédé d'une préposition 'Y'

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.

1.3 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.

Chapter 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

2.3 Extraction des Descripteurs

2.4 Méthodes de Classification

2.4.1 Bayésien Naïf

2.4.2 Hidden Markov Model

2.4.3 Conditional Random Field

Chapter 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

Chapter 4

Discussions

Conclusion

Appendix 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

Appendix B

Algorithmes

Bibliography

[Greenwade, 1993] Greenwade, G. D. (1993). The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351.