LETTRES
SORBONNE
UNIVERSITÉ

Projet Méthodologie

Création des features et évaluation de leur efficacité

GAO SHUAI

Introduction

- But : Créer des « features » et évaluer leur efficacité dans la différenciation des textes de genres différents
- Corpus : Corpus Brown du Natural Language Toolkit
- Outils : Codes en Python fournis en cours et Weka (Waikato Environment for Knowledge Analysis)

Définition du corpus

- Corpus Brown du NLTK
  - Constitution
    - « reviews », « editorial » → « news »
    - « science_fiction », « fiction », « mistery » → « literature »
    - « learned » → « sciences »

LETTRES SORBONNE UNIVERSITÉ

Choix des « features »

```
{
    "verbs_present_tense":
        [
            "be", "am", "are", "is", "have", "has",
            "do", "does", "say", "says", "go", "goes",
            "get", "gets", "make", "makes", "know", "knows",
            "think", "thinks", "take", "takes",
            "see", "sees", "come", "comes", "want", "wants",
            "look", "looks", "use", "uses", "find", "finds",
            "give", "gives", "tell", "tells", "work", "works",
            "call", "calls", "try", "tries", "ask", "asks",
            "need", "needs", "feel", "feels", "become", "becomes"
        ],

    "verbs_past_tense":
        [
            "was", "were", "had",
            "did", "said", "went",
            "got", "made", "knew",
            "thought", "took",
            "saw", "came", "wanted",
            "looked", "used", "found",
            "gave", "told", "worked",
            "called", "tried", "asked",
            "needed", "felt", "became"
        ]
}
```

Tests de classification dans Weka

all_features

	F-Mesure	Rappel	Précision	Instances correctement classifiés
J48	0,686	0,720	0,720	72%
NaiveBayes	0,600	0,620	0,613	62%
RandomForest	0,537	0,560	0,534	56%

verbs_past_tense

	F-Mesure	Rappel	Précision	Instances correctement classifiés
J48	0,686	0,720	0,720	72%
NaiveBayes	0,611	0,640	0,618	64%
RandomForest	0,548	0,560	0,545	56%

verbs_present_tense

	F-Mesure	Rappel	Précision	Instances correctement classifiés
J48	?	0,460	?	46%
NaiveBayes	0,450	0,480	0,481	48%
RandomForest	0,403	0,400	0,433	40%

Discussion des résultats

- Résultats
  - « Features » peu efficaces
    - Confusion de formes des verbes au passé et des participes passés
    - Présence sous-estimés des verbes au présent
- Perspectives
  - Prise en considération des formes composées
  - Élimination des verbes au présent dans « features »

LETTRES
SORBONNE
UNIVERSITÉ

Merci