Classification d'opinions par méthodes symbolique, statistique et hybride

Sigrid Maurel, Paolo Curtoni

AFIA 2007 - DEFT'07

Grenoble, 3 juillet 2007

Introduction

Contexte

- classification d'opinions, présents dans des textes de différents domaines
- corpus : critiques de films, de livres et de jeux vidéo, relectures d'articles scientifiques, textes de débats politiques

CELI France

- entreprise privée spécialisée dans le « *Sentiment Analysis* » et l'« *Opinion Mining* » (analyse des opinions)
- développement de trois méthodes pour classer les textes des différents corpus
  - symbolique
  - statistique
  - hybride

La méthode symbolique

- analyse syntaxique du texte par un analyseur fonctionnel et relationnel
- l'analyse se fait au niveau des phrases
  - découpage du texte en phrases
  - analyse des phrases, extraction d'information (sous forme de relations)
- vérification pour chaque phrase si elle contient des relations de sentiment
- grammaire spéciale pour l'extraction des relations de sentiment (positives, négatives et moyennes)

Grammaire des sentiments

- une grammaire pour l'extraction des relations de sentiments a été développée pour le domaine du tourisme
- elle a été adaptée aux corpus DEFT'07
  - une grammaire spécifique pour chaque corpus
  - ajout de règles pour les sentiments moyens
  - modifications du lexique pour chaque corpus
- pas de grammaire pour le corpus des débats politiques

Les relations syntaxiques

- relations de base : modifieur d'un nom (*une belle maison*) ou d'un verbe (*lire attentivement*)
- relations plus complexes : le sujet d'un verbe (*Pierre fait des courses*), la coréférence (*la ville de Grenoble qui se trouve dans les Alpes*)
- relations de sentiment
  - le sentiment et sa cause (*j'aime beaucoup Grenoble*)
  - la polarité

  $\Rightarrow$ notation : `SENTIMENT_POSITIF(aimer,Grenoble)`

Les relations de sentiment

- pour les sentiments positifs et négatifs calcul à base de mots marqués avec un trait spécial dans le lexique
  - surtout des adjectifs (*magnifique, affreux*) et des verbes (*aimer, regretter*)
  - dans des relations de modifieur, sujet et objet
- pour les sentiments moyens calcul d'après la construction de la phrase
  - coordination d'un sentiment positif et d'un sentiment négatif (*un livre passionnant mais inabouti*)
  - présence de mot-clés comme par exemple *pourtant, malgré* (*Malgré un début superbe...*)
- inversion de la polarité dans le cas d'une négation (*un restaurant pas cher*)

Listes de termes

- création de listes de termes selon le domaine du corpus
- chaque liste contient les mots pour lesquels on souhaite extraire les relations
  - *aVoiraLire* : film, livre, album, ...
  - *jeuxvidéo* : jeu, graphisme, jouabilité, ...
  - *relectures* : article, texte, résultat, ...
- les relations extraites portant sur d'autres mots ne sont pas considérées

$\Rightarrow$ les sentiments dans les résumés du film, livre, etc. ne sont pas pertinents pour le sentiment global du texte

Calcul de l'opinion du texte

- le nombre de sentiments positifs, moyens et négatifs est retenu pour chaque phrase
- à la fin du texte les sentiments sont calculés et mis en relation pour donner un sentiment global du texte entier
- un indice de confiance est ajouté au sentiment global pour la méthode hybride

La méthode statistique

- basée sur des techniques de l'apprentissage automatique
- adaptation à la langue française (n-gram = 12) pour le projet du tourisme
- puis utilisation sur les corpus DEFT'07, en ajoutant une méthode pour les sentiments moyens
- entrainement et classification au niveau des textes entiers

Fonctionnement

- extraction des phrases qui contiennent des sentiments à l'aide de la méthode symbolique
- entrainement des modèles (un pour chaque corpus) sur les extraits des textes
- classification des textes
- calcul d'un indice de confiance pour la méthode hybride

Expérimentations

- avec les textes du corpus *aVoiraLire*
- entrainement du modèle uniquement sur les premières et/ou dernières phrases du texte
- hypothèse : le résumé du film/livre se trouve au milieu du texte, le jugement au début ou à la fin

$\Rightarrow$ meilleurs résultats qu'avec les textes entiers

- abandon de cette technique car difficilement reproductible sur d'autres corpus

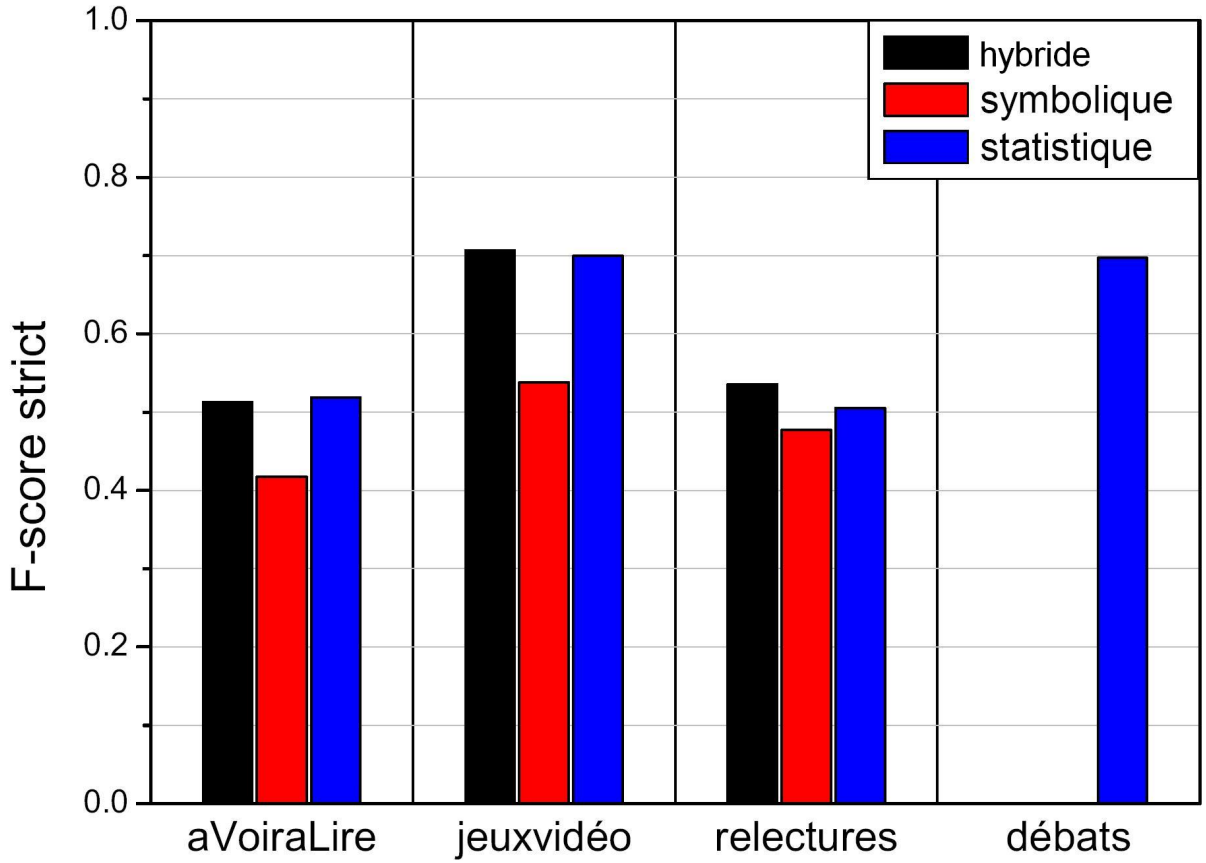
La méthode hybride

- utilisée pour les corpus *aVoiraLire*, *jeuxvidéo* et *relectures d'articles*
- comparaison des résultats des deux méthodes précédentes
- calcul du résultat global d'après les indices de confiance attribués

$\Rightarrow$ correction de l'apprentissage automatique (méthode statistique) possible par configuration manuelle de la grammaire (méthode symbolique) : lexiques et listes de termes adaptés au domaine

Résultats

⇒ meilleurs résultats avec la méthode hybride pour les corpus *jeuxvidéo* et *relectures d'articles*



Conclusion

- développement de grammaires de sentiment pour DEFT'07
- adaptation des méthodes symbolique et statistique du domaine du tourisme aux domaines de DEFT'07
- combinaison des méthodes symbolique et statistique a donné des résultats plus précis que chacune des méthodes employée séparément

$\Rightarrow$ possibilité de garder la robustesse de l'apprentissage automatique et d'orienter le résultat dans la direction souhaitée (p.e. d'une application réelle)

Bibliographie

- AÏT-MOKHTAR S., CHANOD J.-P. & ROUX C. (2001). *A multi-input dependency parser.*
- DINI L. (2002). *Compréhension multilingue et extraction de l'information.*
- DINI L. & MAZZINI G. (2002). *Opinion classification through information extraction.*
- PANG B. & LEE L. (2004). *A sentimental education*: *Sentiment analysis using subjectivity summarization based on minimum cuts*.