

L'analyse des sentiments dans les forums

Sigrid Maurel

INFORSID 2008 - FODOP'08

Fontainebleau, 27 mai 2008

- 1 Introduction
- 2 Méthode symbolique
- 3 Méthode statistique
- 4 Méthode hybride
- 5 Évaluation
- 6 Conclusion

- 1 Introduction
- 2 Méthode symbolique
- 3 Méthode statistique
- 4 Méthode hybride
- 5 Évaluation
- 6 Conclusion

Introduction

Contexte

- classification d'opinions positifs et négatifs, présents dans des textes de différents domaines
- corpus : tourisme, DEFT'07, jeux vidéo et imprimantes

CELI France

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

Difficultés

Les difficultés rencontrées

- langage familier et phonétique typique sur Internet
- fautes d'orthographe nombreuses, absence de ponctuation
- exemple de texte du corpus du *tourisme* :

BaLadeur, posté le 13-10-2006 à 11:23:43:

Je partage l'avis d'Aston sur de nombreux points. Villandry est quelconque mais son jardin transformé en potager géant vaut le détour. Chenonceau est certainement le plus photogénique donc le plus connu et il le mérite largement Si tu recherche la monumentalité comme a Versailles, la magnificence en plus, il faut absolument voir Chambort. Enfin s'il faut ne visiter qu'une ville ce sera Tours.

Corpus

Les corpus utilisés

- suggestions de destinations touristiques dans les différentes régions en France et ailleurs dans le monde
- les corpus de DEFT'07 : critiques de livres et films, tests de jeux vidéo, relectures d'articles scientifiques et notes de débats parlementaires, certains contiennent des sentiments moyens
- solutions de problème pour des jeux vidéo
- conseils d'achat pour des imprimantes

- 1 Introduction
- 2 Méthode symbolique**
- 3 Méthode statistique
- 4 Méthode hybride
- 5 Évaluation
- 6 Conclusion

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 (uniquement pour DEFT'07))

La 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
 - pas de grammaire pour le corpus des débats politiques
- puis adaptée aux corpus des *imprimantes* et des *jeux vidéo*
- modifications du lexique pour chaque corpus

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
- ⇒ 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 (uniquement pour DEFT'07) calcul d'après la construction de la phrase
 - présence de mot-clés comme par exemple *pourtant, malgré*
- inversion de la polarité dans le cas d'une négation (*un restaurant pas cher*)

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

- 1 Introduction
- 2 Méthode symbolique
- 3 Méthode statistique**
- 4 Méthode hybride
- 5 Évaluation
- 6 Conclusion

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
- entraînement et classification au niveau des textes entiers

Fonctionnement

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

Expérimentations


- avec les textes du corpus *aVoiraLire* (critiques de films, livres, ...)
 - entraînement 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
- ⇒ meilleurs résultats qu'avec les textes entiers
- abandon de cette technique car difficilement reproductible sur d'autres corpus

- 1 Introduction
- 2 Méthode symbolique
- 3 Méthode statistique
- 4 Méthode hybride**
- 5 Évaluation
- 6 Conclusion

La méthode hybride

- comparaison des résultats des deux méthodes précédentes
 - normalisation des indices de confiance attribués
 - calcul du résultat global en confrontant les indices
- ⇒ correction de l'apprentissage automatique (méthode statistique) possible par configuration manuelle de la grammaire (méthode symbolique) : lexiques adaptés aux domaines en question

Interface de SYBILLE



filter criteria [\(remove all\)](#)

- **Domaine:** "VITESSE" [\(remove\)](#) [\[add more\]](#)
- **Mots Domaine:** "lent" [\(remove\)](#) [\[add more\]](#)

Order Commands

List View
Data Mining
Map View (Unavailable for Sybille)
Graph View
Timeline View

401 items Hello Velocity World!

sorted by URI [A to Z] < previous 1 ... 32 33 34 35 36 37 **38** 39 40 41 next >

materiel/20070509 [URI]

attitude
negative

Domaine
IMPRESSION
VITESSE
HARDWARE

materiel

Secteur
Brother

forum
forum.hardware

text
NEGATIF ~ lent ~ Brother ~ | Je sais que les Brother bas de gamme sont lentes, mais c'est une question de prix.

Expediteur
linuxafficion

Sujet
Multifonction rapport qualité/prix : la nouvelle canon MP500 ? - Imprimantes - Hardware - Périphériques - FORUM HardWare.fr

[\[external link\]](#)

Show Referers

🔍 Type here to search **1**

🔖 **attitude** **2**

Type here to filter

negative (271)

positive (130)

🔖 **Secteur** **3**

Type here to filter

HP (42)

Epson (36)

Canon (30)

Brother (13)

🔖 **Domaine** **4**

Type here to filter

"HARDWARE" (207)

"IMPRESSION" (170)

"QUALITE" (86)

"IMAGE" (81)

"GRAPHISME" (77)

"SYSTEME" (28)

🔖 **Mots Domaine**

Type here to filter

"imprimante" (139)

S. Maurel (CELI France)

Analyse de sentiments

27/05/2008

19 / 25

Interface de SYBILLE, détail

materiel/20060618 [URI]

attitude
positive

Domaine
IMAGE
IMPRESSION
VITESSE

materiel

Secteur
Epson

forum
forum.hardware

text **5**
POSITIF ~ superbe ~ impression ~ | J'en suis très content,elle est très rapide pour le texte et les impressions photos sont superbes et je la consommation d'encre très raisonnable comparé à mon ancienne epson.

Expediteur
Steph657

Sujet
Canon Pixma 6600D - Qualite PARFAITE !!! - Imprimantes - Hardware - Périphériques - FORUM HardWare.fr

[external link] **6**

Show Referers

- 1 Introduction
- 2 Méthode symbolique
- 3 Méthode statistique
- 4 Méthode hybride
- 5 Évaluation**
- 6 Conclusion

Évaluation des trois méthodes






- évaluation avec les corpus de *DEFT'07*
- le F-Score varie entre 0,51 et 0,71 pour la méthode hybride
- meilleurs résultats avec la méthode hybride pour les corpus *jeuxvidéo* et *relectures d'articles*

- 1 Introduction
- 2 Méthode symbolique
- 3 Méthode statistique
- 4 Méthode hybride
- 5 Évaluation
- 6 Conclusion**

Conclusion

- développement de grammaires de sentiment pour différents domaines (tourisme, jeux vidéo, imprimantes, ...)
 - adaptation des méthodes symbolique et statistique à chaque domaine
 - combinaison des méthodes symbolique et statistique donne des résultats plus précis que chacune des méthodes employée séparément
- ⇒ 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.*
-  MAUREL S., CURTONI P. & DINI L. (2007). *Classification d'opinions par méthodes symbolique, statistique et hybride.* In : Actes de DEFT'07.
-  PANG B. & LEE L. (2004). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.*