LAGRATOUNETTE : classification automatique générique de textes d'opinion

Alejandro Acosta, André Bittar

LATTICE-CNRS (UMR 8094), Université Paris 7
{aacosta,abittar}@linguist.jussieu.fr

Résumé : Nous présentons le bilan de la participation de l'équipe de jeunes chercheurs de l'équipe TALANAdu laboratoire Lattice au 3^e Défi Fouille de Textes (DÉFT'07). Le défi de cette année était de classifier les documents de 4 corpus différents selon l'opinion exprimée par chacun d'entre eux. Cet article présente le travail entrepris par notre équipe - notre méthodologie, les ressources utilisées, les étapes suivies lors du traitement et les résultats obtenus par l'application de notre approche.

Mots-clés : classification de documents, textes d'opinion, chaîne de traitement, Weka

1 Introduction

L'approche proposée utilise les modules d'une chaîne de traitement linguistique pour enrichir les documents des corpus DÉFT'07 d'annotations diverses. Ces annotations, et des dictionnaires lexicales générés automatiquement à partir du corpus d'apprentissage ont été utilisés pour la construction de classifieurs pour chaque classe de corpus. Les classifieurs les plus performants ont été choisis avec une plate-forme de exploration de données et utilisés pour le traitement des corpus de test.

Dans un premier temps, notre équipe de jeunes chercheurs a envisagé une approche qui comportait trois niveaux de paramètres à étudier et utiliser dans la tâche de la classification des textes d'opinion. Ces trois niveaux étaient : un niveaux des statistiques sur le comptage d'annotations linguistiques (par exemple, le nombre d'adjectifs, de pronoms clitiques négatifs, etc.), un niveaux modélisant les lexies isolées qui caractérisent chaque classe d'opinion dans la collection de documents utilisée pour l'apprentissage et, finalement, un niveau comportant de paramètres avec des éléments langagières plus complexes (des collocations, des expressions figées, etc.).

Les deux premiers de ces niveaux constituent l'ensemble de paramètres génériques de l'approche ; les techniques utilisées pour calculer les paramètres de ces niveaux peuvent s'appliquer sans aucune modification à une tâche de classification différente, c'est-à-dire qu'il s'agit de paramètres qui ne dépend pas des détails de la tâche de classification des textes d'opinion et qui caractérisent, tout simplement, des textes en langue naturelle.

Les annotations linguistiques, d'une part, ne dépendent pas du domaine du corpus (du type des textes, de leur contenu, des thèmes qu'ils abordent). Quant au vocabulaire, s'il est étudié strictement en isolation (c'est-à-dire en fonction des occurrences des mots dans un corpus d'apprentissage) et non pas en fonction de, par exemple, un champ lexical spécifique à un domaine, les techniques utilisées pour l'exploiter de manière automatique peuvent aussi s'appliquer à tout autre texte.

En revanche, les paramètres du troisième niveau concernent des éléments qui ne pourraient pas vraiment être définis indépendamment du domaine d'application. Dans le cas de la classification des textes d'opinion, la définition de ces paramètres permettrait une analyse plus riche, mais aussi moins générique, car elle repose sur des caractéristiques propres à l'expression de l'opinion (voire propres à l'opinion dans un type de corpus d'opinion).

Malheureusement, tous les membres de l'équipe initiale n'ont pas pu s'investir jusqu'au bout dans le projet prévu. Par conséquent, nous sommes restés au niveau générique de l'apprentissage et de la classification des documents du corpus. Pour ce faire, nous avons procédé en plusieurs étapes :

1. Filtrage du corpus
2. Pré-traitement
3. Construction de dictionnaires de classification

4. Reconnaissance des lexies de classification
5. Calcul des paramètres de classification
6. Evaluation des modèles de classification
7. Ibid 1, 2, 4 et 5 pour le corpus d'évaluation
8. Classification du corpus d'évaluation avec les modèles choisis

On remarquera que nous n'avons finalement pas utilisé des informations sur des structures du troisième niveau présenté ci-dessus. En effet, nous nous sommes limités à une étude élémentaire des paramètres des deux premiers niveaux. Or, nos trouvons qu'il est tout de même important de présenter nos techniques à la communauté qui à participé au DÉFT'07.

Le reste de ce document est organisé de la manière suivante : dans la section 2 nous présentons les grandes lignes de l'approche générique que nous avons utilisée, dans la section 3 nous présentons les idées concernant le filtrage des corpus, dans la section 4 nous présentons la chaîne de traitement linguistique utilisée pour le pré-traitement des corpus, dans la section 5 nous présentons la technique utilisée pour l'obtention des dictionnaires de classification lexicalisés, dans la section 6, nous présentons la technique utilisée pour calculer les paramètres pour chaque document, dans la section 7 nous présentons la plate-forme d'exploration de données utilisée pour l'évaluation des modèles de classification, dans la section 8 nous parlons des résultats obtenus, et finalement dans la section 9 nous présentons les conclusions de notre expérience.

2 Classification générique

Le point de départ de la classification générique est l'ensemble des idées suivantes : les annotations linguistiques peuvent être utilisées comme paramètres pour la classification de documents en langue naturelle. Par ailleurs, certaines formes lexicales ont une importance plus grande que d'autres lorsqu'il s'agit d'identifier des classes différentes. Finalement, il y a des segments des documents qui sont plus importants pour leur classification.

Nous considérons qu'il n'est pas nécessaire d'utiliser la totalité du texte de chaque document pour le classifier. Dans chaque document il peut y avoir des segments qui ne sont pas très pertinents pour sa classification. Il en va de soi qu'il y a un segment (ou des segments) qui sont plus importants que d'autres pour identifier un texte comme appartenant à une classe particulière. Un des buts de notre approche est donc de choisir les extraits les plus pertinents pour la classification des documents et ignorer le reste de leur contenu.

Dans les segments pertinents, ceux qui résultent d'un filtrage des documents, il existe des marqueurs lexicaux précis qui se distinguent par leur association aux classes prédéfinies. Ainsi, on trouvera, par exemple, que des extraits comme *un très bon film* ou *je m'oppose*, sont (dans des corpus de critiques de films et de débats politiques, respectivement) plus utiles que d'autres pour classifier des documents. Ces marqueurs peuvent être des mots isolées ou des expressions plus longues, composées de plusieurs mots[1].

Par ailleurs, le type de langage utilisé dans chaque classe de document (et pour chaque type de corpus) varie en fonction de sa classe. Le texte d'une relecture d'un article qui rejette ce dernier est souvent critique et par conséquent plus négatif qu'affirmatif, par exemple. Les formes et structures choisies par l'auteur d'un texte ne sont donc pas sans rapport avec le type de texte dont il s'agit. On peut utiliser des annotation linguistiques génériques dans le but de capturer les particularités du langage des classes différentes des documents.

L'approche générique consiste alors à filtrer les documents pour ne garder que les segments (qui risquent d'être) pertinents pour leur classification. Les documents filtrés sont ensuite annotés par des modules génériques d'annotation linguistique, et on détecte aussi les occurrences des lexies avec des distributions intéressantes dans les différentes classes. Ces deux types de données (annotations et lexies) sont utilisées pour décrire chaque document, elles deviennent les paramètres de classification.

Le corpus d'apprentissage est utilisé pour trouver un modèle statistique qui donne de bons résultats avec les paramètres qui peuvent être calculés de manière automatique.

[1] On appelle ici *expression*, de manière très générique, une séquence de mots dans un texte. Nous ne faisons aucune hypothèse sur le statut ou la nature linguistique de ces objets.

3 Filtrage des documents

Le choix des organisateurs du DÉFT'07 de diviser le corpus d'évaluation en 4 classes différents est en fait un premier filtrage qui s'opère sur le corpus d'évaluation (là où la classification des textes d'opinion, dans un sens général, pourrait être définie pour tout type de textes). Le résultat est un ensemble de sous-classes de textes d'opinion. Pour chacune de ces sous-classes de textes d'opinion, un autre niveau de classification a été établi, et c'est à ce niveau que les systèmes participant au DÉFT'07 s'intéressent.

Nous considérons que les documents qui font partie de chaque sous-classe peuvent à leur tour passer par un nouveau filtrage. Le but de ce nouveau filtrage est de repérer les segments qui sont plus pertinents pour la classification de chaque classe, pour pouvoir ignorer les segments qui sont moins orientés vers une ou une autre. Si l'on cible les contenus qui nous intéressent à l'intérieur de chaque document, on simplifie le calcul de paramètres.

Le filtrage que nous avons appliqué aux corpus d'apprentissage et d'évaluation a été très élémentaire. Nous nous sommes basés sur nos intuitions et sur un survol des corpus pour déterminer les segments les plus pertinents pour la classification de chacun.

Ainsi, pour le corpus de critiques de films, livres, spectacles et bandes dessinées nous avons choisi de ne garder que les premières 4 phrases de chaque critique. Pour le corpus de tests de jeux vidéo seul le dernier paragraphe (celui employé comme le résumé de la critique) a été gardé pour la classification. Quant au corpus de relectures d'articles , nous avons gardé aussi les 4 premières phrases. Finalement, pour le corpus de débats parlementaires nous avons chois de garder les premières deux et les dernières deux phrases de chaque document, celles qui correspondent, grosso modo, à l'introduction et la conclusion de la participation d'une personne dans un débat.

Bien qu'assez grossier, ce filtrage est une approximation d'un filtrage qui saurait bien distinguer le contenu pertinent de celui qui l'est moins.

4 Les outils MACAON

Après le filtrage, les documents ont été enrichis avec des annotation linguistiques standards. Ces annotations constituent le pré-traitement du contenu qui est utilisé pour calculer les paramètres de classification pour chaque corpus.

MACAON[2] est une architecture modulaire de traitement automatique de langues. Plusieurs modules de cette architecture sont en cours de développement pour l'annotation des textes en français. Les modules utilisés pour l'annotation des documents des corpus DÉFT'07 ont été tirés de cette collection. Il s'agit des modules qui s'occupent des tâches suivantes :

1. Segmentation en phrases
2. Tokenisation
3. Reconnaissance d'entités nommées
4. Analyse lexicale
5. Etiquettage morpho-syntaxique
6. Analyse morphologique
7. Analyse syntaxique partielle

Ces modules ont été appliqués dans l'ordre d'apparition ci-dessus. L'entrée de cette chaîne de modules était donc le texte filtré de chaque document. La sortie est un document XML structuré et enrichi avec des annotations.

5 Construction des dictionnaires

Un dictionnaire a été construit à partir du corpus d'apprentissage pour chaque classe de texte à évaluer, suivant une même procédure.

Le texte des segments retenus après le filtrage (voir section 3) a été découpé en items lexicaux, c'est-à-dire en séquences de caractères séparés par un espace. Ensuite, pour chaque corpus, nous avons calculé :

[2] http ://code.google.com/p/macaon/

1. Le nombre d'occurrences de chaque item
2. Les items uniques à chaque classe
3. La classe maximisant le nombre d'occurrences de chaque item

Ces données nous ont permis de trier les items *uniques* à chaque classe par le nombre de leurs occurrences et de calculer l'importance des items dont les occurrences étaient *maximales* dans chaque classe. Cette importance (I dans la formule 1) résulte de la magnitude de la différence du nombre d'occurrences de l'item dans la classe dans laquelle il apparaît le plus souvent (i_{max}) et la classe dans laquelle il apparaît le moins souvent (i_{inf} avec $i_{inf} > 0$).

$$I(i) = \frac{i_{max} - i_{inf}}{i_{max}} \quad (1)$$

Etant donnée que chaque type de corpus d'apprentissage comportait un nombre très différent de documents, les comptages des occurrences des items lexicaux ont dû être interprétés pour chaque corpus.

Nous n'avons pas implémenté une méthode automatique de sélection des items de chaque dictionnaire. La dernière étape de leur constitution a donc consisté à décider, après un survol des résultats, quels étaient les seuils permettant de trouver un bon compromis entre la quantité d'entrées et la capacité de classification des entrées. Nous noterons, par exemple, que les mots uniques à une classe particulière, mais qui n'apparaissent qu'une fois dans tout le corpus, sont moins importants pour la classification qu'un mot qui apparaît autant de fois qu'il y a de documents.

Les paramètres à déterminer étaient (pour chaque corpus) les suivants :

1. Le nombre maximal d'items *uniques*
2. Le nombre maximal d'items *maximaux*
3. Pour les items *uniques*, le seuil de pertinence du nombre d'occurrences.
4. Pour les items *maximaux*, le seuil de pertinence du nombre d'occurrences.
5. Pour les items *maximaux*, le seuil d'importance de la déviation de ses occurrences

Le module de repérage d'entités nommées de MACAON a été utilisé pour marquer les occurrences des items du dictionnaire dans les corpus d'apprentissage et d'évaluation.

6 Calcul de paramètres

Pour calculer les différents paramètres utilisés pour la création du modèle de classification nous sommes partis des collections de fichiers XML pré-traités correspondant à chaque corpus.

Le calcul des paramètres a été fait par le programme LAGRATOUNETTE, qui prend en entrée une collection de documents pré-traités et une configuration de paramètres à calculer pour donner, en sortie, un fichier dans le format ARFF utilisé par WEKA[3] pour la création du classifieur ou l'application d'un modèle de classification à un corpus.

La configuration des paramètres de LAGRATOUNETTE se fait avec un fichier qui liste les étiquettes à associer aux paramètres (leurs noms) ainsi que la description des éléments à considérer et le type de calcul à effectuer. Ces calculs peuvent être de simples comptages, des facteurs, des déviations de la moyenne dans le corpus, ou simplement la présence ou l'absence d'un élément.

De cette manière on a calculé le nombre d'occurrences des différents parties du discours ; par exemple, la proportion de groupes nominaux par rapport aux groupes verbaux, de groupes verbaux finis par rapport aux groupes verbaux, etc. ; les déviations dans le nombre de dates ou de formes verbales dans les différents modes et temps ; etc.

Dans la figure 1 nous présentons la séquence de tâches effectuées avec les différents corpus lors de l'évaluation. Dans un premier temps le corpus d'évaluation est filtré et pré-traité avec MACAON (en 1), ensuite, on repère les occurrences des éléments des dictionnaires de classification (en 2). Les documents enrichis avec les annotations linguistiques et celles qui correspondent aux lexies sont utilisés par LAGRATOUNETTE (en 3) pour produire les descriptions des documents (en termes de vecteurs de paramètres). Ces descriptions

[3] Voir section 7 pour la présentation de ce logiciel de gestion de classifieurs.

sont ensuite classifiées suivant un modèle (en 4) et, enfin, les résultats de la classification faite par le classifieur, en format ARFF, sont converties en XML (en 5), selon la structure établie par les organisateurs de DÉFT'07.

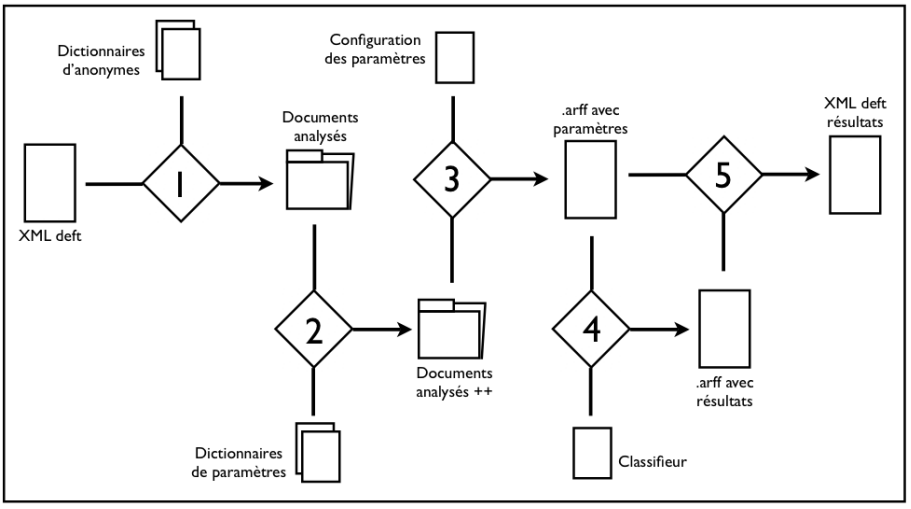

FIG. 1 – La séquence de traitements lors de l'évaluation

7 Evaluation de classifieurs

Weka[4] est un logiciel libre implémenté en Java qui comprend une collection d'algorithmes d'apprentissage (classfieurs) pour des tâches de fouille de données. Il fournit des outils pour le pré-traitement, la classification, la régression, le clustering, des règles d'association et la visualisation des données.

Nous avons utilisé certains des classifieurs pour attribuer une classe, dans notre cas un score soit de 0,1 ou 2, soit de 0 ou 1, à chacun des documents du corpus. Chaque document de corpus est représenté pour WEKA par un ensemble d'attributs (appelé une "instance"). Ces attributs correspondent aux paramètres que nous avons calculés pour chaque document lors du pré-traitement – le nombre d'occurrences de certains mots, étiquettes morphologiques, etc. – ainsi qu'à la note qui lui a été attribuée, dans le cas du corpus d'apprentissage.

Afin de déterminer quel classifieur était le mieux adapté à chaque type de corpus, nous avons effectué des tests préalables, par validation croisée, sur des portions de corpus pré-traités. Les classifeurs qui ont donné les meilleurs résultats pendant ces tests ont été sélectionnés pour l'évaluation.

Un ensemble différent d'attributs a été utilisé selon le type de corpus, les attributs les plus pertinents n'étant pas les mêmes pour tous. Le bilan des paramètres les plus utiles pour la classification est détaillé ci-dessous. Nous présentons ces attributs dans l'ordre de pertinence décroissant pour les 4 corpus.

7.1 Corpus critiques de films, livres, spectacles et bandes dessinées

- **ZERO unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 0. Le graphe ci-dessous représente la distribution des classes de documents selon le nombre d'occurrences de tels mots. L'abscisse représente l'attribut (c'est-à-dire, le nombre de mots de ce type dans un document), et l'ordonné le nombre de documents ayant ce nombre d'occurrences. Chaque barre horizontale est séparée en trois couleurs, chacune d'entre elles représentant une classe (score) où gris foncé = 0, gris moyen = 1 et gris clair = 2. Ainsi, le graphique représente les proportions du total attribuées à chaque classe.

[4]http ://www.cs.waikato.ac.nz/ml/weka/

Pour cet attribut, la figure 2 montre que lorsqu'un document compte un mot de ce type, il y a environ 5 fois plus de chance qu'il soit de score 0 que de score 1, et environ 10 fois plus que de score 2. Si le document compte plus d'un tel mot, il est sûr d'avoir un score de 0.

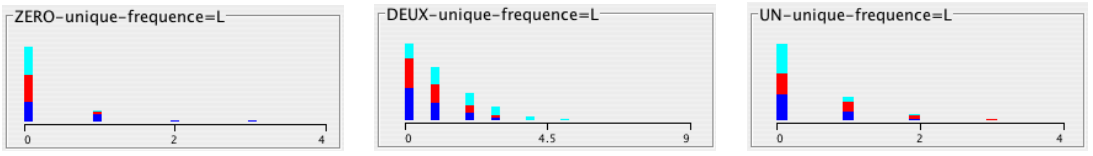

FIG. 2 – 1^{er}, 2^{me} et 3^{me} meilleurs paramètres, corpus critiques de films, livres, spectacles et bandes dessinées

- **DEUX unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 2. Une tendance s'établit lorsqu'un document possède trois mots de cette catégorie, figure 2. Dans ce cas, un score de 2 est environ quatre fois plus probable que 0 ou 1. Avec quatre mots, la probabilité qu'un document ait un score de 2 est environ huit fois plus que pour 1 ou 0. Au-delà de quatre mots, le document est sûr d'avoir un score de 2.

- **UN unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 1. Dans la figure 2, on constate une légère préférence pour un score de 1 lorsqu'un document compte un de ces mots. La distinction devient beaucoup plus marqué pour un document comportant 2 mots de ce type, avec une probabilité de score 1 environ 3 fois plus que pour les autres scores respectivement. A trois mots, il reste une petite probabilité de score 0, mais un score de 1 est massivement plus probable. Au-delà de trois mots, le document est sûr d'avoir un score de 1.

- **comptage dates**
  le nombre d'entités nommées. La figure 3 représente une courbe. On peut constater que dans l'attaque de la courbe (entre 3 et 10 occurrences) la probabilité d'un score 0 est plus importante que pour les autres scores. Au milieu de la courbe, et jusqu'à la chute, les distributions sont relativement proches. A partir de la chute de la courbe, la probabilité d'un score 0 diminue de façon significative et il y a une probabilité plus importante pour un score de 2 (environ deux fois plus probable) que pour un score de 1. Un document comptant un nombre plus élevé d'entités nommées est donc probablement un document de score 2.

- **deviation dates**
  la déviation de la moyenne du nombre d'entités nommées qui sont des dates. La figure 3 représente les tendances inverse de la précédente. En général, pour les valeurs de déviation de la moyenne du nombre de dates entre -70 et -28, la probabilité d'un score de 2 est largement plus élevée. Sur l'attaque de la courbe, les score 2 et 1 sont plus probables que 0. Au milieu de la courbe, et jusqu'à la chute, les distributions sont relativement proches. A partir de la chute de la courbe (une déviation de la moyenne du nombre de dates de 20 à 30), la probabilité d'un score de 0 est plus importante que pour les autres scores.

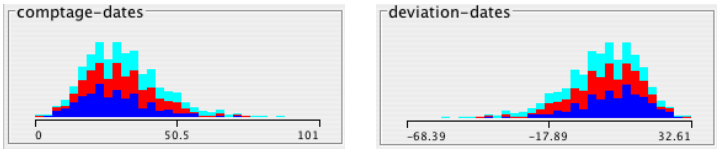

FIG. 3 – 4^{eme} et 5^{me} meilleurs paramètres, corpus critiques de films, livres, spectacles et bandes dessinées

7.2 Corpus tests de jeux vidéo

- **ZERO unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 0. La figure 4 montre qu'avec un mot de ce type, un document est environ quatre fois plus probable d'avoir un score de 0 que chacun des autres notes respectivement. A deux occurrences il y a une probabilité massive d'un score de 0, une minuscule probabilité d'avoir un score de 1, et aucune probabilité de 2. Au-delà, un score de 0 est certain.

- **ZERO unique, fréquence=M**
  Le nombre de mots, ayant une fréquence moyenne, apparaissant uniquement dans les documents de score 0. Dans la figure 4 on remarque qu'avec un mot de ce type, la probabilité d'un score de 0 est presque 1, avec une petite probabilité de 1. Un score de 0 est certain pour un document comptant plus d'une occurrence d'un tel mot.

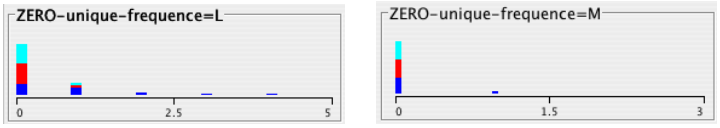

FIG. 4 – 1^{er} et 2^{me} meilleurs paramètres, corpus tests de jeux vidéo

- **DEUX unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 2. La figure 5 montre qu'un score de 2 est environ trois fois plus probable pour un document comptant un mot de ce type et qu'au-delà un score de 2 est une certitude.

- **UN max, déviation=H, comptage=L**
  Ce paramètre représente le nombre de mots apparaissant plus souvent dans les documents de score 1 que dans les autres, dont le nombre d'occurrences n'est pas élevé et dont la déviation de la moyenne d'occurrences dans les classes différentes est importante. La figure 5 montre qu'un document qui ne contient aucun mot de ce type a une forte probabilité d'avoir un score de 0. De 1 à 3 occurrences, les probabilités pour chaque score sont relativement proches, mais à partir de 4 mots la probabilité d'un score de 0 réduit dramatiquement et un score de 1 est environ deux fois plus probable qu'un score de 2.

- **DEUX unique, fréquence=M**
  Le nombre de mots, ayant une fréquence moyenne, apparaissant uniquement dans les documents de score 2. Avec un mot de ce type, la figure 5 montre qu'un score de 2 est environ quatre fois plus probable que les autres scores respectivement. Au-delà, un score de 2 est une certitude.

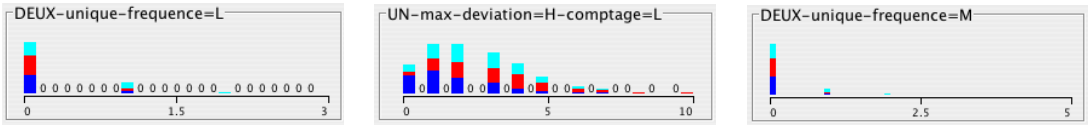

FIG. 5 – 3^{me}, 4^{me} et 5^{me} meilleurs paramètres, corpus tests de jeux vidéo

7.3 Corpus relectures d'articles

- **ZERO unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 0. Comme nous montre la figure 6, un score de 0 est environ deux fois plus probable qu'un score de 1 et environ trois fois plus probable qu'un score de 2 si le document contient un mot

de ce type. A deux et trois occurrences, cette tendance est exagérée. A partir de quatre occurrences, un score de 0 est certain.

- **UN unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 1. Ce graphe, figure 6, représente une courbe descendante, qui montrent une probabilité croissante d'un score de 1 en fonction du nombre d'occurrences d'un mot de ce type. A partir de cinq occurrences, un score de 1 est sûr.

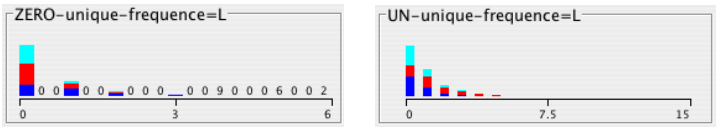

FIG. 6 – 1^{er} et 2^{me} meilleurs paramètres, corpus relectures d'articles

- **ZERO max, déviation=M, comptage=H**
  Le nombre de mots apparaissant plus souvent dans les documents de score 0 que dans les autres, dont le nombre d'occurrences est élevé et dont la déviation de la moyenne d'occurrences dans les classes différentes est assez importante. La figure 7 montrent la même tendance que la précédente, mais pour un score de 0. Plus il y a de mots de ce type dans un document, plus il est probable d'avoir un score de 0.

- **DEUX unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 2. La figure 7 montre encore une courbe descendante. La probabilité d'un score de deux augmente en fonction du nombre d'occurrences des mots de ce type. A partir de quatre mots, un score de 2 est une certitude.

- **DEUX unique, fréquence=M**
  Le nombre de mots, ayant une fréquence moyenne, apparaissant uniquement dans les documents de score 2. Avec un mot de ce type, la figure 7 montre qu'un score de 2 est environ deux fois plus probable qu'un score de 1 ou de 0. A deux occurrences, un score de 2 est quasiment sûr et au-delà devient certain.

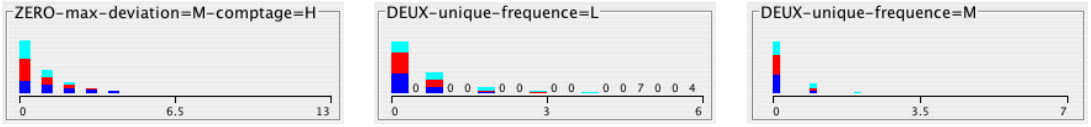

FIG. 7 – 3^{me}, 4^{me} et 5^{me} meilleurs paramètres, corpus relectures d'articles

7.4 Corpus débats parlementaires

- **CONTRE max, déviation=M, comptage=L**
  Ce paramètre représente le nombre de mots apparaissant plus souvent dans les documents classés *contre* que dans ceux classés *pour*, dont le nombre d'occurrences est relativement basse et dont la déviation du nombre d'occurrences dans la classe *pour* est assez importante. La figure 8 représente une courbe descendante où la probabilité d'un score de 0 (vote "contre") augmente en fonction du nombre d'occurrences d'un mot de ce type.

- **POUR unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 1 (vote "pour"). Cette figure, 8, montre qu'avec un mot de ce type, un score de 1 est

environ trois fois plus probable qu'un score de 0. Au-delà, cela devient une certitude.

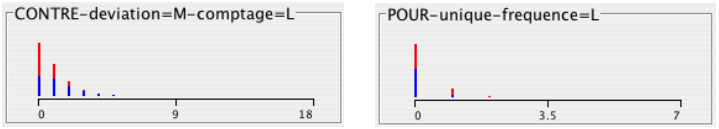

FIG. 8 – 1^{er} et 2^{me} meilleurs paramètres, corpus débats parlementaires

- **CONTRE unique, fréquence=L**
  Le nombre de mots, ayant une fréquence relativement basse, apparaissant uniquement dans les documents de score 0. La figure 9 montre la tendance inverse de la précédente. A une occurrence d'un mot de ce type, un document a une probabilité d'un score de 0 environ trois fois plus élevée que pour un score de 1. Au-delà cela devient une certitude.

- **CONTRE max, déviation=H, comptage=L**
  Ce paramètre représente le nombre de mots apparaissant plus souvent dans les documents classés *contre* que dans ceux classés *pour*, dont le nombre d'occurrences est relativement basse et dont la déviation du nombre d'occurrences dans la classe *pour* est importante. Avec une seule occurrence d'un mot de ce type, la figure 9 montre qu'un score de 0 est environ quatre fois plus probable qu'un score de 1. Au-delà d'une occurrence, un score de 0 est certain.

- **POUR max, déviation=H, comptage=L**
  Ce paramètre représente le nombre de mots apparaissant plus souvent dans les documents classés *pour* que dans ceux classés *contre*, dont le nombre d'occurrences est relativement élevé et dont la déviation du nombre d'occurrences dans la classe *contre* est faible. La tendance ici est similaire. La figure 9 montre qu'avec une occurrence d'un mot de ce type, la probabilité qu'un document ait un score de 1 est environ trois fois plus importante que pour un score de 0. Au-delà d'une occurrence, le document est sûr d'avoir un score de 1.

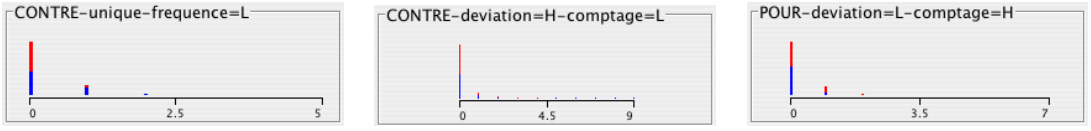

FIG. 9 – 3^{me}, 4^{me} et 5^{me} meilleurs paramètres, corpus débats parlementaires

Lors de notre étude des différents modèles de classification, nous avons aussi évalué la performance des deux différents types de paramètres : les paramètres calculés sur les annotations linguistiques, et ceux issus des dictionnaires. A première vue, on remarque que la présence des paramètres lexicaux est prépondérante dans les listes que l'on vient de présenter. Mais la haute pertinence de ces paramètres n'exclut pas celle des autres.

En effet, quand on regarde les résultats des 5 meilleurs paramètres par eux-mêmes (c'est-à-dire, avec un modèle de classification qui ignore tous les paramètres qui ne sont pas listés dans cette section), on constate une perte importante (d'environ 10 %, pour le corpus d'apprentissage) de performance.

En revanche, la performance ne diminue que de très peu lorsqu'on exclut les paramètres les plus pertinents au moment de la création du modèle de classification. Ces paramètres ne suffisent donc pas, à eux tous seuls, pour modéliser la classification des corpus.

Pour les corpus de critiques de films, livres, spectacles et bandes dessinées , de tests de jeux vidéo et de débats parlementaires, c'est un classifieur J48 qui a donné les meilleurs résultats. Ce type de classifieur est un arbre de décision, une structure simple qui où les noeuds non terminaux représentent des tests sur un ou plusieurs attributs et les noeuds terminaux représentent les décisions prises. Quant au corpus de relectures d'articles, c'est un classifieur Logistic qui a été choisi pour l'évaluation. Ce classifieur implémente la technique de régression logistique, qui prédit les valeurs prises par une variable catégorielle binaire à partir d'une série de variables explicatives continues et/ou binaires.

8 Résultats

Les résultats de l'évaluation envoyés par les examinateurs comportent trois paramètres : la précision, le rappel et un F-score strict. Nous présentons ci-dessous, pour chaque corpus, les résultats sur l'ensemble des soumission ainsi que les résultats de notre équipe avec l'écart de nos résultat vis-à-vis de la moyenne de l'ensemble des participants au DÉFT'07. Les résultats sont présentés dans les tableaux 1 à 4.

Paramètre	Résultats sur l'ensemble	Nos résultats	Ecart de la moyenne
Précision	0.5276 +/- 0.0982	0.3927	0.1349
Rappel	0.4829 +/- 0.0683	0.3920	0.0909
F-score	0.5004 +/- 0.0668	0.3923	0.1081

TAB. 1 – Corpus de critiques de films, livres, spectacles et bandes dessinées

Paramètre	Résultats sur l'ensemble	Nos résultats	Ecart de la moyenne
Précision	0.6925 +/- 0.0996	0.5324	0.1601
Rappel	0.6367 +/- 0.0921	0.5405	0.0962
F-score	0.6604 +/- 0.0864	0.5365	0.1319

TAB. 2 – Corpus de tests de jeux vidéo

Paramètre	Résultats sur l'ensemble	Nos résultats	Ecart de la moyenne
Précision	0.4804 +/- 0.0490	0.4403	0.0401
Rappel	0.4617 +/- 0.0477	0.4348	0.0269
F-score	0.4706 +/- 0.0468	0.4375	0.0331

TAB. 3 – Corpus de relectures d'articles

Comme on a déjà vu dans la section 7, les paramètres les plus pertinents de notre approche ont été les éléments lexicaux dans le corpus d'apprentissage associés à une certaine opinion. Les documents qui ont été classifiés correctement doivent avoir une distribution des lexies semblable à celle qui a produit les dictionnaires générés automatiquement.

En ce qui concerne les documents mal classifiés, on peut supposer que les paramètres génériques n'ont pas suffit. Il reste à voir si l'utilisation de paramètres spécifiques aux textes d'opinion (et même pour les types différents de textes d'opinion) améliore les résultats de manière significative. La réponse se trouve sans doute dans les rapports des équipes qui ont intégré des connaissances spécifiques de ces domaines à la construction de leur modèles de classification.

Il est cependant intéressant de constater que les résultats de notre approche générique ne s'écartent pas trop de la moyenne, surtout dans le cas des corpus de relectures d'articles et de débats parlementaires .

9 Conclusion

Nous avons présenté une approche à la classification des textes d'opinion fondée sur un modèle statistique dont l'apprentissage tient compte de deux types de paramètres : un premier ensemble de paramètres correspond à des statistiques concernant des annotations linguistiques associées aux documents. Un deuxième ensemble correspond à des paramètres issus d'une analyse statistique des items lexicaux avec une incidence importante sur la classification des documents. Ces deux ensembles de paramètres ont été calculés automatiquement ; aucune information lexicale (synonymes, expressions figées, collocations, etc.) extérieur au corpus d'apprentissage n'a été ajoutée avant l'application des modèles de classification au corpus d'évaluation.

Les résultats ne sembleraient pas se trouver parmi les plus performants de la campagne d'évaluation DÉFT'07. Or, compte tenu de l'écart entre les résultats des différentes équipes, il n'est pas sans intérêt de remarquer qu'une approche à la classification qui se contente de filtrer grossièrement les documents et d'utiliser un modèle générique de classification donne déjà des résultats qui ne s'éloignent pas trop de la moyenne.

Paramètre	Résultats sur l'ensemble	Nos résultats	Ecart de la moyenne
Précision	0.6545 +/- 0.0564	0.5820	0.0725
Rappel	0.6298 +/- 0.0645	0.5830	0.0468
F-score	0.6416 +/- 0.0594	0.5825	0.0591

TAB. 4 – Corpus de débats parlementaires

Par ailleurs, nous avons remarqué que les paramètres qui modélisent le type de langage utilisé dans les documents sont, par eux mêmes, assez utiles pour classifier les corpus. En effet, bien que les paramètres issus des dictionnaires d'items lexicaux soient les plus pertinents pour la classification, si l'on ne se sert que de ces paramètres on a une perte de performance importante. Si l'on les exclut, la perte de performance dans les résultats est moindre.

Références

WEISS S. M., INDURKHYA N., ZHANG T. & DAMERAU F. (2005). *Text Mining : Predictive Methods for Analyzing Unstructured Information*. Springer.

WITTEN I. H. & FRANK E. (1999). *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.