Une nouvelle approche pour l'extraction non supervisée de critères

Benjamin Duthil*, François Trousset*
Mathieu Roche**, Michel Plantié*
Gérard Dray*, Jacky Montmain*

*EMA-LGI2P, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France
prénom.nom@mines-ales.fr
**LIRMM Université Montpellier 2, CNRS 5506, 161 Rue Ada, F-34392 Montpellier
prénom.nom@lirmm.fr

Résumé. Récemment de nouvelles techniques regroupées sous le vocable de détection automatique d'opinions (opinion mining) ont fait leur apparition et proposent une évaluation globale d'un document. Ainsi, elles ne permettent pas de mettre en avant le fait que les personnes expriment une opinion très positive du scénario d'un film alors qu'elles trouvent que les acteurs sont médiocres. Dans cet article, nous proposons de caractériser automatiquement les segments de textes relevant d'un critère donné sur un corpus de critiques.

1 Contexte et présentation générale

L'approche décrite dans cet article a pour objectif de réaliser une segmentation de textes selon un domaine et des critères spécifiés par l'utilisateur. Le principe général est le suivant. Dans un premier temps, l'utilisateur spécifie son domaine de recherche (e.g. cinema) et les critères sur lesquels il souhaite faire la segmentation (e.g. acteur, musique, réalisation, ...). Ensuite, pour chaque critère, il spécifie un ensemble de mots germes. Expérimentalement nous considérons que 7 mots germes sont spécifiés pour chacun des critères. Ainsi, dans le cas du critère acteur, nous considérons par exemple les mots germes (en anglais) suivants : acting, actor, casting, character, interpretation, role, star. La seconde étape consiste à utiliser un moteur de recherche (dans nos expérimentations nous avons utilisé google) pour récupérer un ensemble de textes contenant pour des documents de la thématique et pour un critère donné, au moins l'un des mots germes. A l'issue de cette étape nous disposons donc pour chacun des critères, d'un ensemble de 7 classes de documents associés à un mot germe. De la même manière nous interrogeons le moteur de recherche pour obtenir un second ensemble de documents ne contenant aucun des germes pour un critère. L'objectif de ces deux interrogations est la suivante. Dans le premier corpus (i.e. celui contenant des textes associés à un mot germe), nous souhaitons extraire des mots qui sont en corrélation avec des mots germes et ainsi étendre de manière automatique le dictionnaire de mots associés à un critère. Par exemple nous souhaitons pouvoir retrouver via une analyse de ces documents que le mot actor est souvent lié aux mots actress, interpret, player ... L'objectif du second corpus (que nous nommons anti-classe) est au contraire de pouvoir extraire l'ensemble des mots qui ne sont pas reliés au critère

dans une thématique donnée. Dans ce cas, nous souhaitons pouvoir éliminer du bruit au sein de notre dictionnaire. La phase d'apprentissage des mots potentiellement utiles, i.e. qui peuvent être intégrés dans le dictionnaire, est réalisée à l'aide d'un algorithme utilisant une fenêtre glissante autour de tous les mots germes inclus dans le document et en estimant la fréquence des mots les plus associés aux mots germes. Par manque de place, nous ne détaillons pas l'algorithme. A l'issue de cette phase, nous obtenons par critère un ensemble de mots qualifiés de positifs (i.e. ils correspondent au critère) et d'un ensemble de mots négatifs (i.e. ils ne peuvent pas correspondre au critère). Pour chacun de ces mots un score est attribué et tient compte non seulement de la fréquence d'apparition mais également de son apparition dans la classe positive ou dans l'anticlasse.

2 Expérimentations

De manière à évaluer notre proposition, de nombreuses expérimentations ont été réalisées en utilisant comme thématique le cinéma et comme critères acteur et scénario. . Pour chaque critère, nous avons spécifié 7 mots germes et extrait pour chacun de ces derniers 300 documents contenant ce mot germe et 300 documents ne contenant pas ce mot (i.e. l'anticlasse). Les figures 1 et 2 présentent le rappel et la précision pour chacun des critères.

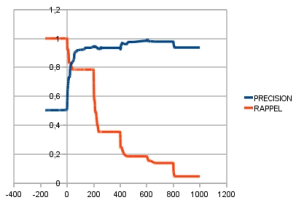

FIG. 1 – *Précision et rappel pour acteur sur une échelle de 0 à 1*

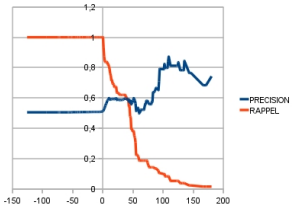

FIG. 2 – *Précision et rappel pour scenario sur une échelle de 0 à 1*

3 Conclusion

Notre approche automatique ne nécessite ni connaissance langagière, ni expertise (sauf la définition des mots germes). Bien que nous soyons conscient de ne pas supprimer la totalité du bruit, nous arrivons quand même à isoler les mots relatifs au critère et identifions grâce à eux des segments de textes pertinents.

Summary

Recently new techniques known as *opinion mining* have emerged. They do not consider that opinions could be expressed through several criteria. Our objective in this paper is to automatically match text segments to given criterias. Conducted experiments on real datasets illustrate the effectiveness of the approach.