Détection de contradiction dans les commentaires

Ismail Badache, Sébastien Fournier, Adrian-Gabriel Chifu

Détection de contradiction dans les commentaires

Ismail Badache — Sébastien Fournier — Adrian-Gabriel Chifu
LSIS UMR 7296 CNRS, Université Aix-Marseille, Marseille, France

RÉSUMÉ. *L'analyse des avis (commentaires) générés par les utilisateurs devient de plus en plus exploitable par une variété d'applications. Elle permet de suivre l'évolution des avis ou d'effectuer des enquêtes sur des produits. La détection d'avis contradictoires autour d'une ressource Web (ex. cours, film, produit, etc.) est une tâche importante pour évaluer cette dernière. Dans cet article, nous nous concentrons sur le problème de détection des contradictions et de la mesure de leur intensité en se basant sur l'analyse du sentiment autour des aspects spécifiques à une ressource (document). Premièrement, nous identifions certains aspects, selon les distributions des termes émotionnels au voisinage des noms les plus fréquents dans l'ensemble des commentaires. Deuxièmement, nous estimons la polarité de chaque segment de commentaire contenant un aspect. Ensuite, nous prenons uniquement les ressources contenant ces aspects avec des polarités opposées (positive, négative). Troisièmement, nous introduisons une mesure de l'intensité de la contradiction basée sur la dispersion conjointe de la polarité et du rating des commentaires contenant les aspects au sein de chaque ressource. Nous évaluons l'efficacité de notre approche sur une collection de MOOC (Massive Open Online Courses) contenant 2244 cours et leurs 73873 commentaires, collectés à partir de Coursera. Nos résultats montrent l'efficacité de l'approche proposée pour capturer les contradictions de manière significative.*

ABSTRACT. *Analysis of opinions (reviews) generated by users becomes increasingly exploited by a variety of applications. It allows to follow the evolution of the opinions or to carry out investigations on products. The detection of contradictory opinions about a Web resource (e.g., courses, movies, products, etc.) is an important task to evaluate the latter. In this paper, we focus on the problem of detecting contradictions based on the sentiment analysis around specific aspects of a resource (document). First, we identify certain aspects, according to the distributions of the emotional terms in the vicinity of the most frequent names in the whole of the reviews. Second, we estimate the polarity of each review segment containing one aspect. Then we take only the resources containing these aspects with opposite polarities (positive, negative). Third, we introduce a measure of the intensity of the contradiction based on the joint dispersion of the polarity and the rating of the reviews containing the aspects within each resource. We evaluate the effectiveness of our approach on the Massive Open Online Courses (MOOC) collection containing 2244 courses and their 73873 reviews, collected from Coursera. Our results show the effectiveness of the proposed approach to capture contradictions significantly.*

MOTS-CLÉS : *Analyse de sentiments, Contenus générés par l'utilisateur, Contradiction.*
KEYWORDS: *Sentiment analysis, User generated content, Contradiction.*

1. Introduction

Au cours des dernières années, l'Internet devient de plus en plus un espace ouvert, social et mobile, où les gens peuvent exprimer leurs opinions en laissant des traces (ex. commentaire, rating, j'aime, etc) sur des ressources web. Il existe de nombreux services qui permettent la génération de ces traces par les utilisateurs, comme les blogs, les wikis, les forums et les réseaux sociaux. Ils représentent tous une source riche d'informations sociales, qui peuvent être analysées et exploitées dans diverses applications et contextes (Badache et Boughanem, 2017). En particulier, l'analyse de sentiments (Htait *et al.*, 2016), par exemple, pour connaître l'attitude d'un client vis-à-vis d'un produit ou de ses caractéristiques, ou pour révéler la réaction des gens à un événement. De tels problèmes nécessitent une analyse rigoureuse des aspects sur lesquels porte le sentiment pour produire un résultat représentatif et ciblé.

Une autre problématique concerne la diversité des opinions sur un sujet donné. Certains travaux l'aborde dans le contexte de différents domaines de recherche, avec une notion différente dans chaque cas. Par exemple, (Wang et Cardie, 2014) visent à identifier des sentiments au niveau d'une phrase exprimée au cours d'une discussion et à les utiliser comme des caractéristiques dans un classifieur qui prédit la dispute dans la discussion. (Qiu *et al.*, 2013) identifient automatiquement les débats entre des utilisateurs à partir du contenu textuel (interactions) dans les forums, en se basant sur des modèles de variables latentes. Il y a eu d'autres travaux dans l'analyse des interactions avec les utilisateurs, par exemple, l'extraction des expressions de type *agreement* et *disagreement* (Mukherjee et Liu, 2012) et d'en déduire les relations de l'utilisateur en regardant leurs échanges textuels (Hassan *et al.*, 2012).

Dans cet article, nous cherchons à comprendre autour de quoi (ex. aspects, sujets) les contradictions peuvent se produire dans les commentaires associés à une ressource web (ex. film, cours, etc) et comment quantifié leur intensité. Nous avons formulés les deux hypothèses suivantes :

Hypothèse 1 : Une contradiction dans des commentaires liés à une ressource donnée signifie des opinions contradictoires exprimées sur un aspect spécifique, qui est une forme de diversité de sentiments autour de l'aspect au sein de la même ressource.

Hypothèse 2 : Un aspect avec un sentiment négatif dans un commentaire avec un rating positif (et inversement) a un impact plus important sur l'intensité de la contradiction qu'un aspect avec un sentiment positif dans un commentaire avec un rating positif.

En outre, une contradiction peut se produire dans un commentaire quand un auteur présente différentes opinions sur le même aspect, ou à travers plusieurs commentaires lorsque différents auteurs expriment des opinions différentes sur le même aspect. Afin de concevoir notre modèle de détection automatique de contradictions, des tâches fondamentales sont effectuées : premièrement, définir automatiquement les aspects caractérisant ces commentaires. Deuxièmement, capturer les opinions opposées autour de chacun de ces aspects à travers un modèle d'analyse de sentiments. Troisièmement, estimer l'intensité de la contradiction au niveau des commentaires pour chaque res-

source, en utilisons une mesure de dispersion. Les questions de recherche abordées dans cet article sont les suivantes :

– Comment identifier une contradiction sur un aspect dans les commentaires ?

– Comment mesurer le degré de contradiction entre les commentaires ?

– Quel est l'impact de la prise en compte conjointe de la polarité et du rating des commentaires sur la mesure de l'intensité de la contradiction ?

L'article est organisé comme suit. La section 2 présente certains travaux connexes. La section 3 détaille notre approche pour la détection des contradictions. L'évaluation expérimentale est présentée dans la section 4. Enfin, la section 5 conclut l'article en annonçant des perspectives.

2. État de l'art

La détection et la mesure de contradiction est un processus complexe qui nécessite l'utilisation de plusieurs méthodes. Plusieurs travaux ont été proposés pour ces méthodes (détection des aspects, analyse de sentiments) mais à notre connaissance, très peu de travaux traitent de la détection et de la mesure de l'intensité de la contradiction. Dans cette section, nous allons brièvement présenter quelques approches de détection de controverses proches de nos travaux puis nous allons présenter les approches liées à la détection des aspects et l'analyse de sentiment, qui sont utiles pour introduire notre approche.

2.1. *Approches de détection de controverses et de faux avis*

Plusieurs travaux sont proches des travaux présentés dans cet article. Ainsi, une des thématiques partageant le plus de point commun avec nos travaux est celle concernant la détection de controverses (*dispute, controversy*). Parmi ces travaux, plusieurs traitent la controverse sur Wikipédia et plus particulièrement dans le cas des commentaires qui entourent les modifications de pages Wikipédia (Wang et Cardie, 2014). D'autres travaux cherchent à détecter les controverses sur des domaines particuliers par exemple dans le cadres de nouvelles (Tsytsarau *et al.*, 2014) ou encore lors de l'analyse de débat (Qiu *et al.*, 2013). D'autres travaux cherchent à être plus génériques en voulant détecter de manière générale la controverse sur le web (Jang et Allan, 2016). Un autre travail plus récent proposé par (Rocktäschel *et al.*, 2015) utilise LSTM (*Long Short-Term Memory*) pour capturer l'itération entre deux phrases (*Entailment, Neutral* ou *Contradiction*). Nos travaux ont aussi une certaine proximité avec les travaux concernant la détection, dans les « *reviews* » de consommateur, de faux avis qui présentent souvent la caractéristique d'être en contradiction (soit positive soit négative) avec les avis majoritaires (Li *et al.*, 2014). Toutefois, à notre connaissance aucun de ces travaux ne cherchent à quantifier l'intensité de la contradiction ou bien de la controverse.

2.2. *Approches de détection d'aspect*

Les premières tentatives de détection d'aspects ont été basées sur l'approche classique d'extraction d'information (IE) en exploitant les phrase nominales fréquentes (Hu et Liu, 2004). De telles approches fonctionnent bien dans la détection des aspects

qui sont sous la forme d'un seul nom, mais sont moins effaces lorsque les aspects sont de faible fréquence. Dans le contexte de la détection d'aspects, bon nombre de travaux utilisent les CRF (*Conditional Random Fields*) ou les HMM (*Hidden Markov Models*). Parmi ces travaux, nous pouvons citer (Hamdan *et al.*, 2015) qui utilise les CRF. D'autres méthodes sont non supervisées et ont prouvé leur efficacité tel que (Titov et McDonald, 2008) qui construisent un modèle thématique à grains multiples (Multi-Grain Topic Model). Nous pouvons aussi citer le modèle HASM (*unsupervised Hierarchical Aspect Sentiment Model*) proposé par (Kim *et al.*, 2013) qui permet de découvrir une structure hiérarchique du sentiment fondée sur les aspects dans les avis en ligne non labellés. Dans nos travaux, nous nous sommes inspirés de la méthode non supervisée développée par (Poria *et al.*, 2014) basées sur l'utilisation de règles d'extraction pour les avis sur les produits.

2.3. *Approches d'analyse de sentiments*

L'analyse du sentiment a fait l'objet de très nombreuses recherches antérieures. Comme dans le cas de la détection d'aspects, les approches supervisées et non supervisées ont chacune leurs solutions. Ainsi, dans les approches non supervisées, nous pouvons citer les approches basées sur les lexiques telles que l'approche développée par (Turney, 2002) ou bien des méthodes basées sur des corpus comme les travaux de (Mohammad *et al.*, 2013). Au rang des approches supervisées, nous pouvons citer (Pang *et al.*, 2002) qui comme nombre de travaux perçoivent la tâche d'analyse de sentiment comme une tâche de classification et utilisent donc des méthodes comme les SVM (*Support Vector Machines*) ou les réseaux bayésiens. D'autres travaux récents sont basés sur les RNN (*Recursive Neural Network*) tels que les travaux de (Socher *et al.*, 2013). Comme le propos de cet article est de vérifier les hypothèses énoncées concernant la détection de contradictions et que l'analyse de sentiment n'est qu'une étape du processus, nous nous sommes inspirés des travaux de (Pang *et al.*, 2002) en utilisant un classifieur Bayésien.

3. Détection de contradiction dans les commentaires

Notre approche de détection de contradiction est basée à la fois sur la détection automatique des aspects au sein des commentaires ainsi que l'analyse de sentiments de ces aspects. Dans cet article, en plus de la détection des contradictions, notre but est également d'estimer l'intensité de ces contradictions entres les commentaires. Mesurer ces contradictions passe à travers une exploitation conjointe et de manière fine de deux dimensions : la polarité autour de l'aspect ainsi que le rating associé au commentaire. Afin de pouvoir prendre en compte ces deux dimensions, nous utilisons une formule de dispersion qui modélise les commentaires comme un nuage de points, dont les coordonnées sont les polarités et les ratings. Notre approche de détection de contradictions pourrait être utiles pour les auteurs des ressources, démontrant quelles informations pourraient nécessiter une vérification plus poussée.

Notre approche consiste à exploiter les contenus sociaux, en particulier les *commentaires* et les *ratings*, comme une source évidente de détection des contradictions dans les commentaires contenants certains aspects spécifiques pour une ressource web.

3.1. *Préliminaires et contexte*

L'information sociale que nous exploitons dans le cadre de notre approche peut être représentée par le triplet <*R, C, N*> où *R, C* et *N* sont des ensembles finis d'instances : *Ressources*, *Commentaires* et *Notes (ratings)*.

Il existe un ensemble $C = \{c_1, c_2, ...c_m\}$ de m commentaires que les utilisateurs peuvent effectuer sur des ressources $R = \{D_1, D_2, ...D_h\}$. Une ressource D peut être un document traditionnel comme une page web ou une ressource web 2.0 comme une vidéo ou toute autre entité similaire. Chaque commentaire généré est suivi par une note attribuée par un utilisateur.

Il existe un ensemble $N = \{Rat_1, Rat_2, ...Rat_m\}$ de m notes (ratings) que les utilisateurs peuvent donner sur les ressources R. Le rating est une note sur une échelle discrète de 1 à une valeur max de 5, où par exemple 3 signifie "moyen" et 5 signifie "excellent".

3.2. *Pré-traitement*

Le pré-traitement est une étape clé pour l'analyse des commentaires (aspects et sentiments). Le module de pré-traitement se compose de trois étapes principales : d'une part, le marquage des termes (identification des noms, verbes, etc), par une analyse syntaxique, au sein des commentaires. Deuxièmement, les noms les plus fréquents dans l'ensemble des commentaires des différents documents sont extraits. Troisièmement, uniquement les noms entourés par des termes émotionnels sont considérés comme des aspects. Nous détaillons ces étapes dans ce qui suit.

3.2.1. *Extraction des aspects*

Définition 1 (Aspect). Dans notre étude, un aspect est une entité nominale très fréquente dans les commentaires étudiés et entourée par des termes émotionnels.

Afin d'extraire les aspects à partir du texte des commentaires, nous avons appliqué les traitements suivants :

1) Calcul fréquentiel des termes constituant le corpus des commentaires,

2) Catégorisation des termes de chaque commentaire en utilisant *Stanford Parser*[1],

3) Sélection des termes ayant la catégorie nominale,

4) Sélection des noms avec des termes émotionnels dans leur voisinages de 5 mots (en utilisant le dictionnaire *SentiWordNet*[2]),

5) Extraction des termes les plus fréquents (utilisés) dans le corpus parmi ceux sélectionner dans l'étape précédente. Ces termes seront considérés comme des aspects.

Exemple : Soit $C = \{c_1, c_2, c_3\}$ un ensemble de 3 commentaires associés à un document D. Nous voulons extraire les aspects à partir de chacun des commentaires en appliquant les étapes décrites ci-dessus.

1. `http://nlp.stanford.edu:8080/parser/`
2. `http://sentiwordnet.isti.cnr.it/`

Nous avons c_1 *="The lecturer was an annoying speaker and very repetitive. I just couldn't listen to him. . . I'm sorry. There was also so much about human development etc that I started to wonder when the info about dogs would start. . . . I found the formatting so different from other courses I've taken, that it was hard to get started and figure things out. Adding to that, was the constant interruption of the "paid certificate" page. If I answer "no" once, please leave me alone ! I also think it's a bit suspect for a prof to be plugging his own book for one of these courses."*

Le tableau 1 récapitule les 5 étapes. Premièrement, nous calculons les fréquences des termes dans l'ensemble des commentaires (à titre d'exemple, les termes "course", "material", "assignments", "content", "lecturer" apparaissent 44219, 3286, 3118, 2947, 2705, respectivement). Deuxièmement, nous étiquetons grammaticalement chaque mots (par exemple, "NN", "NNS" signifient nom en singulier et nom en pluriel, respectivement[3]). Troisièmement, seul les termes de catégorie nominale sont sélectionnés. Quatrièmement, nous gardons uniquement les noms entourés par des termes appartenant au dictionnaire *SentiWordNet* (The *lecturer* was an annoying *speaker* and very repetitive). Enfin, nous considérons comme aspects utiles uniquement les noms qui figurent parmi les noms les plus fréquents dans le corpus des commentaires (l'aspect utile dans ce commentaire est *lecturer*).

Étape	Description
(1)	course : 44219, material : 3286, assignments : 3118, content : 2947, lecturer : 2705, ... terme$_i$
(2)	The/DT **lecturer**/NN was/VBD an/DT annoying/VBG **speaker**/NN and/CC very/RB repetitive/JJ ./. I/PRP just/RB could/MD n't/RB listen/VB to/TO him/PRP .../ : I/PRP 'm/VBP sorry/JJ ./. There/EX was/VBD also/RB so/RB much/JJ about/IN human/JJ **development**/NN **etc**/NN that/IN I/PRP started/VBD to/TO wonder/VB when/WRB the/DT **info**/NN about/IN **dogs**/NNS would/MD start/VB .../ : ./. I/PRP found/VBD the/DT **formatting**/NN so/RB different/JJ from/IN other/JJ **courses**/NNS I/PRP 've/VBP taken/VBN ,/, that/IN it/PRP was/VBD hard/JJ to/TO get/VB started/VBN and/CC figure/VB **things**/NNS out/RP ./. Adding/VBG to/TO that/DT ,/, was/VBD the/DT constant/JJ **interruption**/NN of/IN the/DT "/" paid/VBN **certificate**/NN "/" **page**/NN ./. If/IN I/PRP answer/VBZ "/" no/UH "/" once/RB ,/, please/VB leave/VB me/PRP alone/RB !/. I/PRP also/RB think/VBP it/PRP 's/VBZ a/DT bit/RB suspect/JJ for/IN a/DT **prof**/NN to/TO be/VB plugging/VBG his/PRP$ own/JJ **book**/NN for/IN one/CD of/IN these/DT **courses**/NNS ./.
(3)	lecturer, speaker, development, dogs, formatting, courses, interruption, certificate, page, prof
(4)	lecturer, speaker
(5)	lecturer

Tableau 1 : Les différentes étapes pour extraire les aspects dans un commentaire

Une fois que nous avons défini la liste des aspects utiles qui caractérisent notre collection de données, nous devons estimer la polarité des sentiments autour de ces aspects. La section suivante présente notre modèle d'analyse de sentiments.

3.2.2. *Analyse de sentiment*

Définition 2 (Sentiment). Les sentiments par rapport à un aspect sont un nombre réel dans la plage [-1, 1] qui indique la polarité de l'opinion exprimée dans le commentaire. Les valeurs négatives et positives représentent respectivement des opinions négatives et positives.

Nous utilisons un modèle simple de classification des sentiments supervisé basé sur l'algorithme Naïve Bayes. Naïve Bayes est un modèle probabiliste qui donne de bons résultats lors de la classification des sentiments et prend généralement moins de

3. `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

temps pour la phase d'entraînement par rapport à des modèles comme SVM (*Support Vector Machines*). Un autre avantage de Naïve Bayes est qu'il ne nécessite qu'une petite quantité de données d'entraînement pour estimer les paramètres nécessaires à la classification (Ayetiran et Adeyemo, 2012). Le modèle Naïve Bayes implique une hypothèse d'indépendance conditionnelle simplifiée. Cela est donné par une classe (dans notre cas la polarité : positive ou négative), les mots sont conditionnellement indépendants les uns des autres.

Dans notre cas, la probabilité de vraisemblance maximale d'un mot appartenant à une polarité donnée est calculée par la formule suivante :

$$P(x_i|Pol) = \frac{\textit{Nombre de } x_i \textit{ dans les commentaires de la polarité Pol}}{\textit{Nombre total des mots dans les commentaires de la polarité Pol}} \quad [1]$$

Définition 3 (commentaire-aspect). Il existe un ensemble $CA = \{ca_1, ca_2, ...ca_n\}$ de n commentaires-aspects que les utilisateurs ont exprimé par rapport à un aspect donné sur les ressources R. Chaque commentaire-aspect ca représente un extrait (une fenêtre de 5 mots avant et après l'aspect) du commentaire initial c.

Selon la règle de Bayes, la probabilité d'un commentaire-aspect particulier appartenant à une polarité Pol est donnée par la formule suivante :

$$P(Pol|ca) = \frac{P(ca|Pol) \cdot P(Pol)}{P(ca)} \quad [2]$$

$$P(Pol|ca) = \frac{(\prod P(x_i|Pol)) \cdot P(Pol)}{P(ca)} \quad [3]$$

Où :

– les valeurs de la variable $Pol \in \{Positive, Negative\}$.

– ca représente l'extrait du commentaire contenant l'aspect.

– x_i représente les mots du commentaire-aspect ca.

Si le classifieur rencontre un mot qui n'a pas été observé dans l'ensemble d'apprentissage, la probabilité des deux classes deviendrait nulle et il serait impossible de comparer. Ce problème peut être résolu par le lissage laplacien comme suit :

$$P(x_i|Pol) = \frac{Count(x_i) + k}{(k+1) \cdot \textit{(Nombre de mots de la polarité } Pol\textit{)}} \quad [4]$$

Habituellement, $k = 1$. Puisque *Bernoulli Naïve Bayes* est utilisé, le nombre total de mots dans une classe (polarité) est calculé par rapport à chaque commentaire-aspect, dont ce dernier est réduit à un ensemble de mots uniques sans doublons.

3.2.3. *Traitement des négations*

Un problème majeur rencontré lors de la tâche de classification des sentiments est celui de traiter les négations. Puisque nous utilisons chaque mot comme caractéristique, le mot *"great"* dans l'expression *"not great"* contribuera au sentiment positif plutôt qu'au sentiment négatif (car le *"not"* n'est pas pris en compte).

Afin de remédier à ce problème, nous avons conçu un algorithme simple pour traiter les négations. Nous avons utilisé une représentation alternative des formes négatives comme le montre (Das et Chen, 2001). Notre algorithme utilise une variable d'état pour stocker l'état de négation. Il transforme un mot précédé d'un *"no", "not"* ou *"n't"* en *"not_"+mot*. Chaque fois que la variable d'état de négation est vérifiée, les mots lus sont traités comme *"not_"+mot*. La variable d'état est réinitialisée lorsqu'un signe de ponctuation (" ?., ! : ;") est rencontré ou lorsqu'il y a une double négation.

Puisque le nombre de formes négatives peut ne pas être suffisant pour un apprentissage correcte. Nous avons abordé ce problème en ajoutant des formes négatives à la classe opposée avec les formes normales de toutes les caractéristiques pendant la phase d'entraînement. C'est-à-dire nous essayons d'équilibrer les formes négatives vis-à-vis des formes normales des mêmes mots. Il s'agit de s'assurer que le nombre de formes *"not_"* soient suffisants pour la classification.

3.2.4. *Traitement d'intensificateurs et d'adverbes*

En général, l'information sur le sentiment est véhiculée par des adjectifs ou par certaines combinaisons d'adjectifs avec d'autres mots comme *"very", "surely"*. Ces informations peuvent être capturées en ajoutant des fonctions traitant des paires de mots (*bigrams*), ou même des triplets de mots (*trigrams*). Des mots comme *"very"* ou *"absolutely"* ne fournissent pas de sentiment sur eux-mêmes, mais des phrases comme *"very bad"* ou *"absolutely recommended"* impactent la polarité du texte négativement ou positivement. En incluant ces intensificateurs et adverbes (en forme de *bigrams* et *trigrams*), nous avons pu capturer cette information au sujet des adjectifs. L'utilisation de *bigrams* et *trigrams* nécessite une quantité importante de données dans l'ensemble d'apprentissage, mais ce n'est pas un problème car notre ensemble d'entraînement avait 50.000 commentaires. Toutefois, les données peuvent ne pas être suffisantes pour utiliser 4-grams, cela peut conduire à un sur-apprentissage (*overfitting*) de l'ensemble d'entraînement.

3.3. *Mesure de contradiction*

Nous rappelons que la principale problématique que nous voulons résoudre dans cet article est la détection efficace d'avis contradictoires dans les commentaires (sur des aspects spécifiques), ainsi que leur intensité.

Un commentaire sur une ressource donnée couvre un aspect ou plusieurs sur un domaine général (par exemple, cours, films, médias). Pour chacun des aspects abordés dans un commentaire, nous souhaitons identifier le sentiment exprimé. Dans cette étude, nous nous limitons à identifier et à enregistrer la polarité de ces sentiments comme il a été décris précédemment. Dans la suite, nous nous référons à ces polarités pour détecter les commentaires contradictoires et estimer l'intensité de leur contradiction en se basant à la fois sur le rating et le sentiment de la portion du commentaire contenant l'aspect.

Définition 4 (Contradiction). Il y a une contradiction sur un aspect, entre deux extraits de commentaires contenant cet aspect, $ca_1, ca_2 \in D$ dans un document D, où *Polarité*$(ca_1) \cap$ *Polarité*$(ca_2) = \phi$, lorsque les avis autour de l'aspect sont opposés. *Polarité*(ca_i) représente la fonction qui retourne la polarité (positive, négative) de ca_i.

Pour pouvoir identifier des opinions contradictoires, nous devons définir une mesure de contradiction. Dans notre cas, nous estimons le degré de contradiction entre les commentaires abordant un aspect donné en fonction de deux dimensions : la polarité du commentaire-aspect Pol ainsi que son rating Rat. Nous supposons que plus la distance est élevée entre ces valeurs par rapport à chaque commentaire-aspect ca_i du même document D, plus le degré de contradiction est important.

Soit $ca(Pol, Rat)$ un point du plan. Nous construisons l'indicateur de dispersion par rapport au centroïde $ca_{centroïde}$ que nous notons $Disp(ca^{Pol}_{Rat}, D)$:

$$Disp(ca^{Pol}_{Rat}, D) = \frac{1}{n}\sum_{i=1}^{n} Distance(Pol_i, Rat_i) \quad [5]$$

avec :

$$Distance(Pol_i, Rat_i) = \sqrt{(Pol_i - \overline{Pol})^2 + (Rat_i - \overline{Rat})^2} \quad [6]$$

$Distance(Pol_i, Rat_i)$ est la distance du point ca_i du nuage au point centroïde $ca_{centroïde}$, tandis que n est le nombre de points (commentaires-aspect) du nuage. Nous sommes amenés ici à additionner Pol_i^2 et Rat_i^2, il est donc essentiel que ces deux grandeurs soient normalisées. La polarité Pol_i est une probabilité, mais les valeurs des ratings Rat_i doivent être normalisées comme suit : $Rat_i = \frac{Rat_i - 3}{2}$.

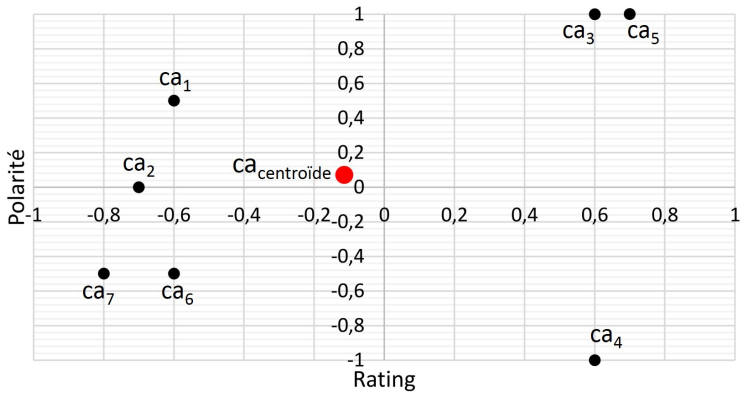

Figure 1 : Dispersion des commentaires-aspect ca_i par rapport au centroïde ca

En affectant chaque point du nuage ayant la même masse $1/n$, l'indicateur $Disp(ca^{Pol}_{Rat}, D)$ représente l'inertie du nuage par rapport au centroïde $ca_{centroïde}$.

On vérifie les propriétés qui confèrent à $Disp$ le statut d'indicateur de dispersion :

– $Disp$ est positif ou nul ; $Disp = 0$ signifie que tous les points du nuage sont confondus en $ca_{centroïde}$ (absence de dispersion).

– $Disp$ croit lorsqu'on éloigne de $ca_{centroïde}$ un point du nuage (c'est-à-dire lorsqu'on augmente la dispersion).

Les coordonnées (Pol, Rat) du centroïde $ca_{centroïde}$ peuvent être calculées de deux manières différentes. Une manière simple consiste à calculer la moyenne des points, dans ce cas le centroïde $ca_{centroïde}$ correspond au point moyen des coordonnées $ca_i(Pol_i, Rat_i)$. Une autre manière plus fine consiste à pondérer cette moyenne par la différence en valeur absolue entre les deux valeurs des coordonnées (dimensions : polarité et rating).

3.3.1. *Centroïde basé sur la moyenne des polarité et des ratings*

Soit la série statistique à deux variables (dimensions), Pol et Rat, dont les valeurs sont des couples (Pol_i, Rat_i). On appelle centroïde (point moyen de la série) basé sur la moyenne des polarité et des ratings le point $ca_{centroïde}$ de coordonnées :

$$\overline{Pol} = \frac{Pol_1 + Pol_2 + ... + Pol_n}{n} \quad [7]$$

$$\overline{Rat} = \frac{Rat_1 + Rat_2 + ... + Rat_n}{n} \quad [8]$$

3.3.2. *Centroïde basé sur la moyenne pondérée des polarité et des ratings*

Soit la série statistique à deux variables, Pol et Rat, dont les valeurs sont des couples (Pol_i, Rat_i). On appelle centroïde (point moyen pondéré de la série) basé sur la moyenne pondérée des polarités et des ratings le point $ca_{centrode}$ de coordonnées :

$$\overline{Pol} = \frac{c_1 \cdot Pol_1 + c_2 \cdot Pol_2 + ... + c_n \cdot Pol_n}{n} \quad [9]$$

$$\overline{Rat} = \frac{c_1 \cdot Rat_1 + c_2 \cdot Rat_2 + ... + c_n \cdot Rat_n}{n} \quad [10]$$

avec :

$$c_i = \frac{|Rat_i - Pol_i|}{2n} \quad [11]$$

Dans cette représentation vectorielle bidimensionnelle, avec les dimensions polarité et le rating, nous proposons l'hypothèse qu'un point dans cet espace a plus d'importance si les valeurs de chacune des deux dimensions sont le plus écartées. Autrement dit, nous croyons qu'un aspect négatif dans un commentaire avec un rating élevé a plus de poids et inversement. Par conséquence, nous avons calculé un coefficient d'importance pour chaque point de l'espace. Ce coefficient est basé sur la différence en valeur absolue entre les deux valeurs des dimensions. La valeur absolue assure que la valeur du coefficient est positive et la division par $2n$ représente une normalisation par la valeur maximale de la différence en valeur absolue ($max(|rating - polarit|) = 2$) et par le nombre de points dans l'espace n qui est le nombre de commentaires-aspect ca. Par exemple, pour une polarité de -1 et un rating de 1, la valeur du coefficient sera de $1/n$ ($|-1-1|/2n = 2/2n = 1/n$). Par contre, pour une polarité de 1 et un rating de 1, la valeur du coefficient sera 0 ($|1-1|/2n = 0$).

4. Cadre expérimental

Afin de valider notre approche, nous avons réalisé une série d'expériences sur les commentaires issus du site de *Coursera*[4]. Notre objectif principal dans ces expériences est d'évaluer l'impact de la prise en compte de l'analyse de sentiment et le rating sur la détection de contradictions dans les commentaires autour de certains aspects spécifiques identifiés automatiquement. Nous évaluons également, l'impact du centroïde moyenné et pondéré sur la mesure de l'intensité des contradictions.

4.1. *Description de la collection de test*

A notre connaissance, il n'existe pas à ce jour de collection de test standard pour l'évaluation de l'efficacité des systèmes de détection de contradictions dans les commentaires. Dans le but d'expérimenter l'efficacité de notre approche, nous avons collecté 2244 ressources en anglais extraits du site "coursera.org" via son API[5]. Chaque ressource décrit un cours et est représentée par un ensemble de métadonnées. Pour chaque cours, nous avons collecté également ses commentaires et ses ratings via le *parsing* des pages web des cours (voir les statistiques sur le tableau 2).

Champ	Nombre Total
Cours	2244
Cours notés	1115
Commentaires	73873
Ratings	298326
Commentaires avec Ratings "1"	1705
Commentaires avec Ratings "2"	1443
Commentaires avec Ratings "3"	3302
Commentaires avec Ratings "4"	12202
Commentaires avec Ratings "5"	55221

Tableau 2 : Les chiffres des données de la collection de Coursera.org

Nous avons pu capturer automatiquement 22 aspects utiles à partir de l'ensemble des commentaires (voir Tableau 3). Pour obtenir des jugements de contradiction et de sentiments pour un aspect donné : 1) nous avons demandé à trois utilisateurs (2 femmes et 1 homme) d'évaluer la classe de sentiment pour chaque commentaires-aspect de 10 cours ; 2) trois autres utilisateurs (3 hommes) ont évalué le degré de contradiction entre les commentaires-aspect. En moyenne 6 commentaires-aspect par cours sont jugés manuellement pour chaque aspect (totalement : 1320 commentaires-aspect de 220 cours, c'est-à-dire 10 cours pour chaque aspect). Nous notons que chaque aspect a été jugé par 3 utilisateurs.

Pour évaluer les sentiments et les contradictions dans les commentaires-aspect de chaque cours, nous utilisons une échelle de notation de 3 points pour les sentiments : (*Negative, Neutral, Positive*) ; et une echelle de 5 points pour les contradictions : *Not Contradictory, Very Low, Low, Strong* et *Very Strong* (voir la figure 2).

4. `https://www.coursera.org/`

5. `https://building.coursera.org/app-platform/catalog`

Aspects	#Rat 1	#Rat 2	#Rat 3	#Rat 4	#Rat 5	#Négatif	#Positif	#Comment	#Cours
Assignment	204	208	333	840	1726	1057	1763	2384	186
Content	176	179	341	676	1641	505	1496	1883	207
Exercise	29	46	94	290	693	195	531	673	58
Information	100	123	238	523	1389	299	1165	1359	143
Instructor	129	106	122	302	1514	295	1107	1322	140
Knowledge	74	72	121	400	1604	905	791	1243	178
Lecture	185	206	290	613	1762	763	1508	1988	208
Lecturer	32	41	48	85	461	55	193	236	39
Lesson	40	59	75	224	712	187	420	554	84
Material	191	203	328	722	2234	784	1693	2254	237
Method	19	23	40	125	404	53	187	224	31
Presentation	46	50	75	142	413	93	196	274	54
Professor	76	74	129	452	3001	331	2234	2369	151
Quality	55	53	51	110	372	113	170	262	54
Question	94	98	172	284	356	311	289	502	104
Quizz	151	155	221	401	581	481	475	824	128
Slide	56	64	81	121	115	131	102	192	47
Speaker	17	15	34	70	170	34	72	103	24
Student	140	105	171	383	1035	519	709	1066	172
Teacher	62	46	82	293	2180	248	1481	1642	119
Topic	67	89	176	437	1154	236	951	1066	130
Video	228	238	356	707	1614	941	1421	2058	245

Tableau 3 : Statistiques sur les aspects issus des commentaires de Coursera.org

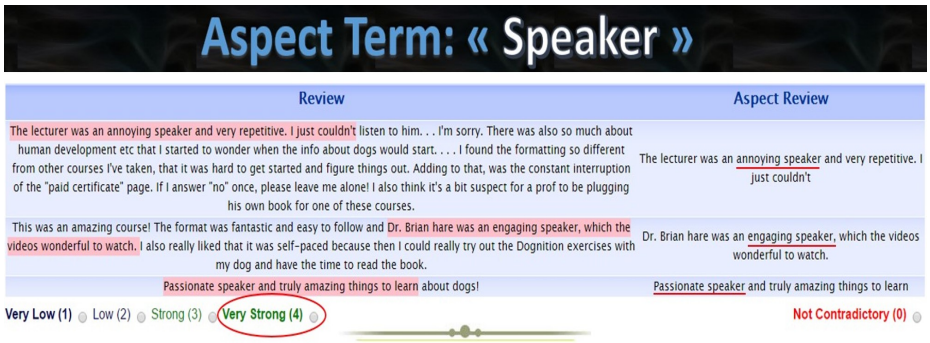

Figure 2 : Interface du système d'évaluation

Nous avons analysé le degré d'accord entre les évaluateurs des contradictions pour chaque aspect avec la mesure Kappa Cohen k (Cohen, 1960). Cet indicateur prend en compte la proportion d'accord entre les évaluateurs et la proportion de l'accord attendu entre les évaluateurs par hasard. La mesure de Kappa est égale à 1 si les évaluateurs sont complètement d'accord, 0 s'ils ne sont d'accord que par hasard. k est négatif si l'accord entre évaluateurs est pire que l'aléatoire.

La figure 3 montre la distribution de la mesure kappa pour chaque aspect. Nous constatons que la mesure de l'accord varie de 0.41 à 0.88. La mesure moyenne d'accord entre les évaluateurs est de 68%, ce qui correspond à un accord fort.

Concernant l'analyse du degré d'accord entre les évaluateurs des sentiments, nous avons trouvé un accord de Kappa $k = 0.76$, qui correspond aussi à un accord fort.

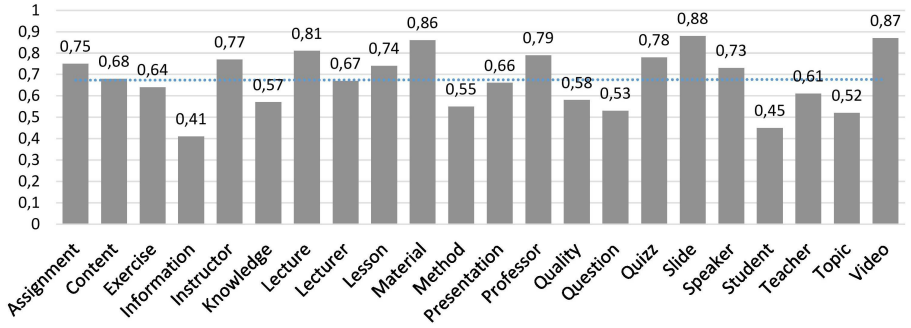

Figure 3 : Répartition de la mesure Kappa k par aspect. < 0 désaccord, $0.0 - 0.2$ accord très faible, $0.21 - 0.40$ accord faible, $0,41 - 0,6$ accord modéré, $0,61 - 0,80$ accord fort, $0,81 - 1$ accord parfait.

4.2. *Résultats et discussions*

Pour évaluer la performance de notre approche, nous avons effectué une analyse de corrélation qui est une des mesures officielles des tâches de SemEval[6]. Nous avons utilisé les coefficients de corrélation de *Pearson* et *Spearman* (Bolboaca et Jantschi, 2006), entre les jugements de contradiction donnés par les évaluateurs et les résultats obtenus par notre approche.

Remarques : Premièrement, notre analyseur de sentiments prend comme ensemble d'apprentissage 50.000 commentaires de films IMDb[7], et comme ensemble de test nos commentaires-aspect de Coursera. Deuxièmement, nous notons que notre système d'analyse de sentiments fourni une exactitude de 79% (c-à-d le taux d'erreur est de 21%) selon l'étude de corrélation. Troisièmement, nous considérons que les jugements des évaluateurs sur les sentiments, représentent une exactitude de 100%.

Les figures 4 et 5 présentent les valeurs de corrélations obtenues par les deux configurations présentées dans les sections 3.3.1 (centroïde basé sur la moyenne des ratings et des polarités) et 3.3.2 (centroïde basé sur la moyenne pondérée des ratings et des polarités). Les résultats de détection de contradictions sont obtenus en se basant sur notre modèle d'analyse de sentiments (voir section 3.2.2) pour la figure 4, alors que pour la figure 5 nous nous sommes basés sur les jugements manuels de sentiments.

a) Centroïde basé sur la moyenne des dimensions. Les résultats montrent que la mesure de dispersion basée sur un centroïde moyenné apporte une corrélation positive avec les jugements de contradiction, Pearson : 0.45, 0.68 et Spearman : 0.42, 0.65 (voir configuration (1), figures 4 et 5). En effet, plus les polarités entre les commentaires-aspect s'opposent plus les points du nuage se dispersent par rapport au centroïde, d'où l'intensité de contradiction augmente. En outre, les résultats obtenus en utilisant les

6. `http://alt.qcri.org/semeval2016/task7/`
7. `http://ai.stanford.edu/~amaas/data/sentiment/`

jugements de sentiments manuels (configuration (1) figure 5) surpassent ceux obtenus en utilisant notre modèle d'analyse de sentiments (configuration (1) figure 4) avec un taux approximatif de 50%. Par conséquent, perdre 21% de précision en sentiment implique une perte d'environ 50% dans la détection de contradiction.

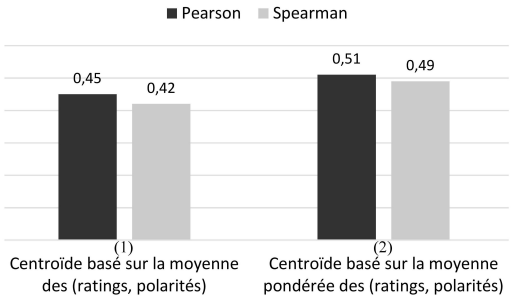

Figure 4 : Corrélation entre les jugements de contradiction et les résultats de notre approche (avec un taux d'erreur de 21% dans la détection de sentiments)

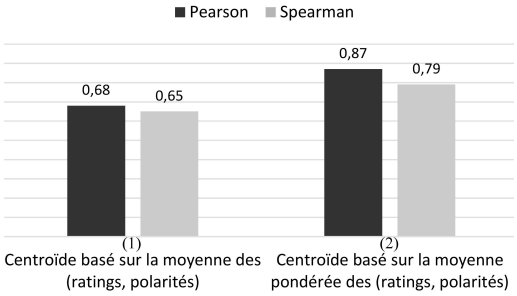

Figure 5 : Corrélation entre les jugements de contradiction et les résultats de notre approche (avec un taux d'erreur de 0% dans la détection de sentiments)

b) Centroïde basé sur la moyenne pondérée des dimensions. Nous remarquons également avec la configuration (2) selon les figures 4 et 5, que les résultats sont positifs (Pearson : 0.51, 0.87 et Spearman : 0.49, 0.79). Les résultats obtenus en prenant compte le coefficient d'importance pour chaque point de l'espace (commentaire-aspect *ca*) sont meilleurs par rapport à ceux obtenus lorsque ce coefficient est ignoré. Ces améliorations sont de 16% (Spearman) en utilisant notre modèle de sentiments (voir figure 4) et de 22% (Spearman) en utilisant les jugements de sentiments manuels (voir figure 5). En effet, plus les valeurs de rating et de polarité pour chaque commentaire-aspect sont divergentes, plus l'impact est important dans l'intensité de contradiction. Nous constatons aussi que les résultats de la configuration (2) présentés dans la figure 5 sont beaucoup plus meilleurs (Spearman : 0.87) que ceux présentés dans la figure 4 (Spearman : 0.51). Ceci revient à la précision des polarités de sentiments. Par conséquent, le modèle d'analyse de sentiments représente un facteur important qui impacte la détection et la mesure des contradictions.

Enfin, les figures 4 et 5 montrent que parmi toutes nos expériences, les meilleurs résultats sont obtenus par la configuration (2) qui prend en considération le coefficient d'importance. La formule de dispersion mesurant l'intensité de contradiction devient plus efficace quand elle est combinée avec un modèle d'analyse de sentiments efficace, ce qui mène à une amélioration significative des résultats. Des analyses de sentiments un peu plus approfondies sont nécessaires pour répondre cette question.

5. Conclusion

Dans cet article, nous avons introduis une approche de détection des contradictions, en attirant l'attention sur des aspects dans lesquels les utilisateurs ont des avis contradictoires. Notre approche est basée sur l'extraction des aspects à partir des commentaires et leur analyse de sentiments. De plus, nous décrivons une mesure qui permet d'estimer l'intensité de contradiction en fonction de la dispersion conjointe de la polarité et du rating des commentaires-aspects par rapport à un centroïde. Les coordonnées du centroïde (Pol, Rat) peuvent être calculées soit par la moyenne arithmétique des coordonnées (Pol_i, Rat_i) des points correspondants aux commentaires-aspect ; ou par la moyenne arithmétique pondérée par la différence en valeur absolue entre les deux valeurs des coordonnées de chaque point du nuage. Nous évaluons expérimentalement l'approche proposée en utilisant la collection de données de Coursera. Les résultats obtenus montrent l'efficacité de notre approche. De plus les meilleurs résultats ont été obtenus grâce à la méthode basée sur la moyenne pondérée ce qui vérifie la seconde hypothèse que nous avons émis à savoir qu'un avis négatif sur un aspect dans un commentaire avec un rating fortement positif (et inversement) a un impact plus élevé sur la perception des utilisateur sur l'intensité de la contradiction.

Le problème potentiel de cette approche c'est qu'elle dépend de la qualité du modèle d'analyse de sentiment. Comme l'ensemble d'apprentissage (commentaires IMDb) est relativement différent de l'ensemble de test (commentaires Coursera), si un mot dans l'ensemble d'apprentissage apparaît uniquement dans une classe et n'apparaît dans aucune autre classe, dans ce cas, le classifieur classera toujours le texte à cette classe particulière. De plus, nous ne gérons pas la détection des phrases à lesquelles appartient un aspect. Nous utilisons une fenêtre prédéfinie de 5 mots avant et après l'aspect. D'autres expérimentations à plus grande échelle sur d'autres types de collections sont également envisagées. Ceci étant même avec ces éléments simples, les premiers résultats obtenus nous encouragent à investir davantage cette piste.

Remerciements

Ce travail a été réalisé grâce au soutien du projet A*MIDEX (ANR-11-IDEX-0001-02), financé par le programme gouvernemental "Investissements d'Avenir", géré par l'Agence Nationale de la Recherche.

6. Bibliographie

Ayetiran E. F., Adeyemo A. B., « A data mining-based response model for target selection in direct marketing », *Journal of Information Technology and Computer Science*, 2012.

Badache I., Boughanem M., « Fresh and Diverse Social Signals : any impacts on search ? », *ACM SIGIR on Conference on Human Information Interaction and Retrieval*, 2017.

Bolboaca S. D., Jantschi L., « Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds », *LJS*, 2006.

Cohen J., « A coefficient of agreement for nominal scales », *Educational and psychological measurement*, vol. 20, n° 1, p. 37-46, 1960.

Das S. R., Chen M. Y., « Yahoo ! for Amazon : Sentiment parsing from small talk on theWeb », *Meeting of the European Finance Association*, 2001.

Hamdan H., Bellot P., Bechet F., « Lsislif : CRF and logistic regression for opinion target extraction and sentiment polarity analysis », *SemEval*, p. 753–758, 2015.

Hassan A., Abu-Jbara A., Radev D., « Detecting subgroups in online discussions by modeling positive and negative relations among participants », *EMNLP*, p. 59-70, 2012.

Htait A., Fournier S., Bellot P., « LSIS at SemEval-2016 Task 7 : Using Web Search Engines for English and Arabic Unsupervised Sentiment Intensity Prediction », *International Workshop on Semantic Evaluation*, p. 469-473, 2016.

Hu M., Liu B., « Mining and summarizing customer reviews », *ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 168-177, 2004.

Jang M., Allan J., « Improving Automated Controversy Detection on the Web », *ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 865-868, 2016.

Kim S., Zhang J., Chen Z., Oh A. H., Liu S., « A Hierarchical Aspect-Sentiment Model for Online Reviews. », *AAAI Conference on Artificial Intelligence (AAAI-13)*, 2013.

Li J., Ott M., Cardie C., Hovy E. H., « Towards a General Rule for Identifying Deceptive Opinion Spam. », *ACL*, p. 1566-1576, 2014.

Mohammad S. M., Kiritchenko S., Zhu X., « NRC-Canada : Building the state-of-the-art in sentiment analysis of tweets », *SemEval*, 2013.

Mukherjee A., Liu B., « Mining contentions from discussions and debates », *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

Pang B., Lee L., Vaithyanathan S., « Thumbs up ? : sentiment classification using machine learning techniques », *EMNLP*, p. 79-86, 2002.

Poria S., Cambria E., Ku L.-W., Gui C., Gelbukh A., « A rule-based approach to aspect extraction from product reviews », *workshop on SocialNLP*, p. 28-37, 2014.

Qiu M., Yang L., Jiang J., « Modeling interaction features for debate side clustering », *ACM Conference on information & knowledge management*, p. 873-878, 2013.

Rocktäschel T., Grefenstette E., Hermann K. M., Kočiskỳ T., Blunsom P., « Reasoning about entailment with neural attention », 2015.

Socher R., Perelygin A., Wu J. Y., Chuang J., Manning C. D., Ng A. Y., Potts C., « Recursive deep models for semantic compositionality over a sentiment treebank », *EMNLP*, 2013.

Titov I., McDonald R., « Modeling online reviews with multi-grain topic models », *Proceedings of the 17th international conference on World Wide Web*, ACM, p. 111-120, 2008.

Tsytsarau M., Palpanas T., Castellanos M., « Dynamics of News Events and Social Media Reaction », *KDD*, p. 901-910, 2014.

Turney P. D., « Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews », *ACL*, p. 417-424, 2002.

Wang L., Cardie C., « A Piece of My Mind : A Sentiment Analysis Approach for Online Dispute Detection », *Annual Meeting of the Association for Computational Linguistics*, 2014.