

Extraction d'opinions pour l'analyse multicritère à partir de corpus oraux transcrits : État de l'art

Emmanuelle Kelodjoue

Laboratoire Informatique de Grenoble, Université Grenoble Alpes

700 Avenue centrale, 38401 Saint Martin d'Hères, France

emmanuelle.kelodjoue-nguemegne@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article présente une revue de la littérature dans les domaines de l'analyse de sentiments et du résumé automatique appliqués aux corpus oraux transcrits. Peu de travaux ont été réalisés dans ces deux domaines avec des corpus oraux transcrits. C'est pourquoi nous présentons ci-dessous les tendances générales dans ces deux domaines et nous nous concentrons ensuite sur les systèmes développés pour le traitement de ce type de données. Les méthodes supervisées pour l'analyse de sentiments et les méthodes extractives pour le résumé automatique sont actuellement dominantes dans le développement des systèmes automatiques pour le traitement des corpus oraux transcrits. Le présent article est une vue d'ensemble de l'état de l'art en analyse de sentiments et du résumé automatique appliqués à des corpus oraux transcrits pour une aide à l'analyse des verbatim.

ABSTRACT

Opinion extraction for multi-criteria analysis of transcribed oral corpora.

This article presents a review of the literature in the areas of sentiment analysis and automatic summarization applied to transcribed oral corpora. Little work has been done in these two areas with transcribed oral corpora. For this reason, we present general trends in these two domains and then focus on the systems developed for processing these type of data. Supervised methods for sentiment analysis and extractive methods for automatic summarization are currently dominant in the development of automatic systems for the processing of transcribed oral corporuses. This article is an overview of the state of the art in sentiment analysis and automatic summarization applied to transcribed oral corpora in order to propose a review system of verbatim.

MOTS-CLÉS : Analyse de sentiments, résumé automatique, corpus oraux transcrits, subjectivité.

KEYWORDS: Sentiment analysis, automatic summarization, spoken transcribed corporuses, subjectivity.

1 Introduction

L'analyse des verbatim a considérablement évolué ces dernières années avec le développement d'Internet et la multiplication des contenus en ligne. Elle permet de caractériser à l'aide de diverses méthodes (analyse lexicale, sémantique, fouille de données, etc.) les paroles échangées dans le cadre d'une conversation. Poland (1995) définit le terme "verbatim" comme la « reproduction mot pour mot de données verbales ». Les mots écrits sont la réplique exacte des paroles enregistrées. Cette reproduction inclut donc toutes les répétitions, les interjections, le bégaiement et les marqueurs

discursifs tels que *heu, hein, ben, etc.* Il existe des campagnes d'évaluation proposant des tâches en analyse de sentiments (DEFT (Défi Fouille Texte) (Nzali *et al.*, 2017), SemEval (Pontiki *et al.*, 2016)), et des tâches de résumé automatique TAC (Text Analysis Conference) (Collins *et al.*, 2017), mais les données mises à disposition pour évaluer les systèmes construits ont été essentiellement des données textuelles (tweets, critiques cinématographiques, messages de forums, etc.). De même, les visées de ces tâches en analyse de sentiments portaient essentiellement sur l'étude de la polarité des sentiments et/ou des "opinions". Nos travaux ont pour objectif de proposer un système d'aide à l'analyse des verbatim d'opinion issus d'entretiens semi-directifs. Par verbatim d'opinion, nous entendons tout énoncé indiquant les opinions et/ou les ressentis des personnes interrogées au sujet d'une thématique présentée dans le cadre d'une conversation. Pour retrouver ces opinions, il est donc important de capturer dans ces transcriptions d'entretiens le contenu subjectif y étant énoncé et le catégoriser. Compte-tenu de notre objectif, nous proposons d'explorer deux domaines de recherche : l'analyse de sentiments et le résumé automatique. Le premier utilise les méthodes en traitement du langage naturel pour automatiser l'extraction et la classification des sentiments dans les documents. Le second permet d'extraire l'information et de condenser l'information pertinente contenue dans les textes pour en produire une synthèse.

Cet article est organisé comme suit : la section 2 présente le problème de l'analyse de sentiments en caractérisant ses approches, la section 3 est dédiée aux types et approches de résumé automatique et dans la section 4, nous présentons une synthèse des deux états de l'art et nous concluons par nos perspectives.

Contribution : Un état de l'art de l'analyse de sentiments et du résumé automatique appliqués aux corpus oraux transcrits pour une aide à l'analyse des verbatim.

2 Analyse de sentiments

L'analyse de sentiments est un sous-domaine de la fouille de textes et se distingue de ses méthodes classiques (qui visent à retrouver des faits) car elle s'intéresse à la détection d'opinions et à leur caractérisation à l'aide d'attributs tels que la polarité, la subjectivité, l'émotion ou encore l'intensité. Lorsque l'on parle d'analyse de sentiments, il s'agit essentiellement d'une tâche de classification. Cette tâche peut suivre trois directions principales. La première consiste à déterminer si un texte est subjectif ou objectif, la deuxième s'attelle à classifier les textes subjectifs en positifs ou négatifs ; enfin, la troisième essaie de déterminer le degré de positivité ou de négativité des textes (Wiebe *et al.*, 2001). Cette tâche peut être appliquée au niveau du texte, ou encore à des niveaux intermédiaires comme le paragraphe, la phrase, ou le mot.

2.1 Approches génériques

Il existe deux approches pour traiter automatiquement le problème de l'analyse d'opinions : les approches fondées sur les lexiques et les approches fondées sur l'apprentissage automatique. Quelques travaux ont également combiné ces deux méthodes.

2.1.1 Approches fondées sur les lexiques

Elles utilisent un lexique de mots, généralement des adjectifs auxquels une polarité est attribuée. Elles ne nécessitent ni de données étiquetées, ni de phases d'entraînement. Les lexiques utilisés dans ces approches peuvent être construits manuellement (Abdulla *et al.*, 2013), semi-automatiquement (Abdaoui *et al.*, 2017) ou en utilisant de grand corpus ou des thésaurus (Goyal & Daumé III, 2011). Pour chaque polarité, un score est assigné par différents algorithmes.

Il existe des lexiques de sentiments qui attribuent un score de -1 aux mots négatifs et +1 aux mots positifs (Mulki *et al.*, 2017). La polarité d'un texte est calculée en cumulant la somme des scores des termes négatifs et positifs. La polarité totale est déterminée en observant le signe de la valeur obtenue. On peut également trouver des dictionnaires où chaque mot est associé à une valence ou intensité pouvant aller de +1 à +5 (Abdulla *et al.*, 2013). Les mots avec une polarité d'intensité +5 sont considérés comme très positifs par rapport aux mots ayant une intensité de +1. On a également des dictionnaires dans lesquels chaque entrée a deux scores de sentiments (positif et négatif) (Mulki *et al.*, 2017). La polarité est calculée en additionnant les scores positifs et les scores négatifs dans le texte. L'orientation du texte est déterminée en observant le score le plus élevé.

Bien qu'il existe plusieurs lexiques de sentiments pour la langue anglaise, nous n'avons connaissance que d'un très petit nombre de lexiques disponibles pour le français :

- Feel¹ (Abdaoui *et al.*, 2017) est un lexique obtenu en traduisant le lexique anglais NRC EmoLex. Il est composé de 14 129 mots. Une polarité (positive ou négative) est attribuée à chaque mot ainsi qu'une émotion parmi les six émotions suivantes : joie, colère, peur, surprise, dégoût et tristesse.
- JeuxDeMots² est un lexique polarisé obtenu à travers le jeu en ligne *likeit* (Lafourcade *et al.*, 2015). Le jeu consiste à assigner à chaque mot une des trois polarités disponibles. Un nombre de votes pour chaque polarité est alors généré à partir des réponses des joueurs. De plus, une liste de 21 émotions à attribuer aux mots est également proposée. Les joueurs peuvent également rajouter une nouvelle émotion lorsqu'ils jugent qu'elle est manquante dans la liste préétablie. Cette ressource contient à ce jour 824 434 mots.
- Polarimots³ (Gala & Brun, 2012) est un lexique de polarité construit en utilisant la deuxième version de POLYMOTS⁴, une ressource lexicale qui est composée de 19 009 mots répartis en 2 069 familles morpho-phonologiques. La ressource Polarimots est composée de 7 483 mots dont des noms, verbes, adjectifs et adverbes.
- Le lexique Affect (Augustyn *et al.*, 2006) contient environ 1 300 mots français ainsi que leurs polarités respectives. Construit automatiquement, il inclut également 45 catégories hiérarchiques émotionnelles.

Bien que les méthodes fondées sur l'utilisation de lexiques sont moins complexes à mettre en œuvre et peu coûteuses en temps de calcul, elles sont très limitées. En effet, elles ne tiennent pas compte du contexte (Ding *et al.*, 2008) dans lequel les mots sont employés. Ainsi un mot peut avoir un sens positif dans un certain contexte et un sens négatif dans un autre. C'est le cas, par exemple, du mot «simple» qui est négatif lorsqu'il est associé à «personne» et positif lorsqu'il est associé à «solution».

1. <http://www.lirmm.fr/abdaoui/FEEL>

2. <http://www.jeuxdemots.org/jdm-accueil.php>

3. <http://polarimots.lif.univ-mrs.fr/>

4. <http://polymots.lif.univ-mrs.fr/v2/>

2.1.2 Approches fondées sur l'apprentissage automatique

On les regroupe en deux catégories : les méthodes supervisées et les méthodes non-supervisées. Dans cette partie, nous nous sommes davantage focalisés sur les méthodes supervisées.

Les méthodes supervisées sont les plus utilisées pour la tâche de classification. Deux ensembles de données sont nécessaires : les données d'entraînement et les données de test. Les données d'entraînement sont utilisées par un classifieur pour apprendre à caractériser et différencier les données. Les données de test servent quant à elles à évaluer la performance du classifieur sur la tâche de classification. Pour que ces méthodes soient efficaces, elles doivent être apprises sur de grands corpus. Des prétraitements linguistiques (stemming, suppression des mots-vides et ponctuations, lemmatisation, etc.) peuvent également être appliqués aux corpus pour réduire leur densité. Une fois le corpus d'entraînement obtenu, il est non seulement important de sélectionner la technique d'apprentissage à utiliser mais aussi de déterminer les descripteurs. Le choix des descripteurs permet d'avoir une représentation structurelle du document et d'améliorer la performance des systèmes. Parmi les descripteurs les plus utilisés, on retrouve : les n-grammes ou uni-grammes, les catégories morpho-syntaxiques (Parts of speech), les mots d'opinion (opinion words), et la négation. Les algorithmes d'apprentissage les plus utilisés pour la tâche de classification d'opinions sont les modèles Naïfs de Bayes (NB), les machines à vecteurs de support (SVM) ou encore les réseaux de neurones. Ces derniers ont récemment montré de très bonnes performances dans les tâches de détection d'opinions (Gaillat *et al.*, 2018; Moore & Rayson, 2017).

D'après les recherches que nous avons effectuées, l'approche par apprentissage profond est celle qui a été retenue dans la majorité des travaux récents en détection d'opinions dans les corpus écrits tels que les réseaux sociaux (Cliche, 2017; Ali *et al.*, 2018) et les avis de consommateurs (Kooli & Pigneul, 2018; Barhoumi *et al.*, 2018).

Les travaux ayant recours aux méthodes supervisées se fondent uniquement sur un corpus annoté pour effectuer la tâche de classification. Dans la majorité des cas, la construction du corpus se fait manuellement, ce qui est très consommateur en temps et en effort humain.

Les méthodes non-supervisées utilisent moins d'informations que les méthodes supervisées. Elles ne nécessitent pas de données annotées. Ces méthodes ont pour but de regrouper les données par paquets ou *clusters* (Liu & Zhang, 2012). Ces clusters sont ensuite triés en fonction de leur importance et de leur pertinence. Contrairement à l'apprentissage supervisé, l'apprentissage non-supervisé exige beaucoup moins travail et de données.

2.1.3 Approches hybrides

Plusieurs travaux ont exploré les approches hybrides en combinant des méthodes de classification à des techniques utilisées pour la construction de lexiques de sentiments. Dans cette partie, nous n'en présentons que quelques-uns.

Hedar & Doss (2013) proposent un classifieur fondé sur un algorithme SVM pour classer des commentaires en langue arabe sur Facebook. En outre, un lexique de sentiments d'expressions argotiques et idiomatiques (SSWIL⁵) est construit à partir du vocabulaire utilisé dans les commentaires. Leurs expériences montrent que l'utilisation du lexique SSWIL combiné au classifieur améliore en termes

5. Slang Sentimental Words and Idioms Lexicon

de précision la tâche de classification avec 86.86 % au lieu de 75.35 % avec un lexique classique d'opinion. Maurel *et al.* (2008) confirment que la combinaison d'une méthode statistique et d'une méthode symbolique donne des résultats plus précis que les deux méthodes utilisées séparément. Ils proposent le système hybride *SYBILLE* qui combine les deux. La méthode symbolique s'appuie sur un parseur qui analyse le texte et extrait pour chaque phrase les relations de sentiments à l'aide d'un lexique de mots polarisés. Une relation de sentiments a deux arguments : le premier est une expression linguistique qui véhicule un sentiment et le deuxième est l'objet ou la cause du sentiment. Pour la phrase *Je déteste les bonbons*, la relation sera SENTIMENT_POSITIF(détester, bonbons). Une grammaire est appliquée à l'ensemble des textes contenant des opinions et dans un deuxième temps à chaque texte selon la thématique générale qu'il contient. S'agissant de la méthode statistique, ils utilisent une technique d'apprentissage automatique fondée sur la représentation n-gramme de caractères ($n = 12$). Pour la tâche d'entraînement, les algorithmes SVM et NB sont utilisés. Le système *SYBILLE* prend donc comme entrée la sortie des autres systèmes et calcule selon un indice de confiance pour chaque résultat, une moyenne qui sera traduite en score positif ou négatif. La méthode statistique permet d'obtenir les polarités des messages et la grammaire qui a été configurée peut être modifiée ou améliorée pour l'obtention de meilleurs résultats. L'approche prend la forme d'un cycle itératif ou les résultats sont constamment améliorés.

2.2 Travaux en analyse d'opinions sur des corpus oraux transcrits

Dans la section précédente, nous avons abordé les types d'approches en analyse d'opinion. Dans ce qui suit, nous décrivons les différentes méthodes adoptées pour le traitement de corpus oraux transcrits.

Dans le domaine politique, plusieurs travaux ont abordé le problème de la classification d'opinions à des débats du Congrès américain. Thomas *et al.* (2006) utilisent un modèle de classification supervisée (SVM) pour déterminer si les segments de discours transcrits sont en faveur ou non d'un projet de loi. Ils soulignent que la spécificité du matériau à analyser implique l'existence de plusieurs relations (accord ou désaccord) entre segments de discours appartenant au même débat, qui peuvent être exploitées pour améliorer la classification. De même, ils notent que dans le cadre de débats parlementaires, les orateurs auront tendance à passer plus de temps à défendre leurs arguments en faveur ou en désaccord d'une loi qu'à exprimer leur opinion.

Pour une tâche connexe, Abercrombie & Batista-Navarro (2018) introduisent deux modèles de classification des transcriptions de débats parlementaires (HanDeSet⁶) : le premier effectue une classification de chaque unité du corpus et, dans le deuxième, les discours donnés en réponse à des motions classées positives ou négatives sont classifiés séparément. Pour comparer la performance des deux modèles de classification, ils utilisent deux classifieurs : un classifieur SVM et un perceptron multicouche. Ils évaluent également l'utilisation de représentations n-grammes (uni, bi-, et trigrammes) et d'informations contextuelles (identifiant du débat, l'affiliation politique de l'orateur et le nom du parti soumettant la législation). Les résultats montrent que le deuxième modèle d'analyse associé aux informations contextuelles améliore les résultats pour les deux classifieurs. Ils remarquent également que les informations contextuelles peuvent être très prédictives des résultats du vote des membres.

Pérez-Rosas & Mihalcea (2013) proposent de construire un classifieur de sentiments s'appuyant uniquement sur les descripteurs linguistiques présents dans des transcriptions d'avis oraux. Ils essaient

6. <https://data.mendeley.com/datasets/xsvp45cbt4/2>

de déterminer s'il existe une différence significative entre l'analyse de sentiment des avis oraux et des avis écrits. Leurs résultats suggèrent que les avis oraux amènent à une performance égale ou inférieure du classifieur comparée aux avis écrits. Ceci s'explique par le fait que les personnes ont tendance à moins exprimer leurs avis si elles se savent observées ce qui est le cas dans les vidéos d'avis. Ils soulignent par ailleurs une perte de précision lorsque les transcriptions manuelles sont remplacées par les transcriptions automatiques.

Cailliau & Cavet (2010) proposent un modèle de modélisation du déroulement émotionnel de conversations téléphoniques en polarité positive et négative. Premièrement, ils évaluent la dégradation d'un système de détection d'entités sur des transcriptions automatiques. Pour pallier ce problème, ils proposent ensuite un modèle qui s'appuie sur la détection de cinq types d'entités évaluatives. Les entités évaluatives correspondent à des mots (adjectif, verbe, nom) ou expressions indiquant les opinions des interlocuteurs tout au long de la conversation. Cinq des types d'entités qu'ils utilisent proviennent des 12 modalités proposées par Charaudeau (2001). À chaque type d'entité, une émotion est également ajoutée. Les entités sont également classées en trois polarités et un score d'intensité leur est attribué. Les transcriptions sont découpées en tours de parole et chaque tour de parole reçoit un score positif et négatif en fonction de la somme des poids d'intensité des entités détectées. La somme des poids d'intensité des entités neutres est alors ajoutée au score le plus élevé des polarités positives ou négatives.

Camelin *et al.* (2006) discutent l'impact des différents modèles de langage et de segmentation automatique sur la tâche de classification appliquée à des messages oraux issus d'un service client. Ils utilisent une méthode de classification reposant sur un algorithme SVM. Deux modèles de langage spécifiques aux opinions et de segmentation ont été développés pour les transcriptions automatiques.

- Le premier modèle (REC01) est de type bigramme et est construit sur la base d'un lexique réduit. Les segments pour ce modèle sont obtenus grâce à une segmentation automatique utilisant un automate pour détecter les pauses réalisées par les locuteurs.
- Le deuxième modèle a pour objectif de modéliser uniquement les portions de messages porteuses d'opinions. Un sous-corpus est extrait pour chaque étiquette qui regroupe les segments appartenant à cette étiquette. Pour chaque étiquette et sous-corpus associé, un sous-modèle bigram est calculé. Un autre modèle englobant a aussi été déterminé pour les étiquettes seules. Les parties ne contenant aucune expression d'opinion sont quant à elles modélisées par une boucle de phonèmes en contexte. Le tout est compilé dans un système nommé REC02. Les segments sont eux obtenus à chaque changement du modèle de langage.

Pour la transcription manuelle, les segments sont obtenus manuellement. Deux expériences sont menées en parallèle et deux types de représentation sont choisies (sac de mots et *seeds*). Les mots *seeds* se réfèrent à un lexique contenant un ensemble de mots susceptibles d'exprimer une polarité positive ou négative. Le lexique a été constitué manuellement et chaque mot lemmatisé afin de généraliser la liste. Pour chaque expérience, les modèles sont comparés avec une référence manuelle (REF). La première expérience concerne *la classification des messages sans segmentation* préalable : Les modèles REC01 et REC02 perdent respectivement 6.5 et 7.7 points de précision par rapport à la référence manuelle lorsque le paramètre choisi est le mot. Le rappel lui décroît de 10 points pour le REC01 alors que le REC02 permet de pallier cette perte avec moins de 3 points. Le recours aux *seeds* permet d'améliorer la F-mesure qui passe de 36.5% et 41.6% à 44.4% et 45.4% respectivement pour les deux modèles. La deuxième expérience concerne *la classification avec segmentation* : les résultats en termes de F-mesure sont plus importants que dans la première expérience quelle que soit la représentation choisie. De même, les résultats concernant la classification et la segmentation effectuées sur le corpus REF montrent une amélioration de 20 points pour la F-mesure. L'introduction

de la segmentation dans la construction des modèles de classification et la représentation des messages permet d'améliorer nettement les premiers résultats.

Parmi les corpus oraux transcrits mis à disposition pour la communauté scientifique et utilisés pour la détection d'opinions, on peut citer le corpus AMI (McCowan *et al.*, 2005) qui a été également utilisé pour la tâche de résumé extractif et abstraktif, le corpus Switchboard (Godfrey *et al.*, 1992), et le corpus RATP-DECODA (Bechet *et al.*, 2012).

D'après les travaux mentionnés ci-dessus, l'analyse de sentiments dans les corpus oraux transcrits traite de plusieurs aspects :

- La robustesse des systèmes de reconnaissance vocale
- La pertinence des modèles d'opinion
- L'adaptation ou non des méthodes standards de l'écrit à l'oral
- La perte de précision des systèmes de détection d'opinions dans les transcriptions automatiques

Les approches par apprentissage supervisé, notamment les classifieurs, SVM, sont celles qui ont été retenues pour la plupart des travaux. Ces méthodes nécessitent un corpus d'entraînement et de test. Le système apprend à caractériser les données à partir de données déjà annotées pour ensuite les généraliser à des données non annotées. Bien que certains comme Cailliau & Cavet (2010) relèvent la présence de traits linguistiques (les marqueurs de discours, les disfluences, les répétitions, etc.) spécifiques au matériau oral ou encore proposent comme Abercrombie & Batista-Navarro (2018) de tenir compte des informations contextuelles du discours, la majorité des travaux abordent la tâche d'analyse de sentiments sur des données orales comme sur des données écrites. Les conclusions des travaux montrent clairement que le choix des descripteurs joue un rôle crucial dans les résultats obtenus.

3 Le résumé automatique

La section précédente a introduit les approches en analyse de sentiments et les méthodes retenues pour le traitement de corpus oraux transcrits. Dans cette section, nous nous intéressons à la problématique du résumé automatique en décrivant ces approches.

3.1 Les types de résumé automatique

Plusieurs travaux en TAL se sont concentrés sur la tâche du résumé automatique. Les premiers travaux étaient ceux de Luhn (1950). On peut également citer les travaux de Nenkova *et al.* (2011) qui ont proposé une classification des différents types de résumés : le résumé informatif, le résumé indicatif, le résumé orienté, le résumé critique, le résumé générique, le résumé de mise à jour. Le résumé informatif présente un ensemble d'informations renseignant sur le contenu d'un texte et tend à conserver l'organisation générale du texte original ainsi que les principales thématiques. Le résumé indicatif doit pouvoir fournir au lecteur les informations essentielles pour l'aider à choisir de consulter ou non le document source. Pour le résumé orienté, l'objectif est de ne résumer que les informations du document d'entrée qui répondent à une requête spécifique de l'utilisateur. Le résumé critique est un résumé contenant non seulement les informations pertinentes du texte mais également l'analyse critique du document original. Le résumé générique est supposé éviter au lecteur de lire le texte source en présentant de manière succincte la thématique générale du texte. Le résumé de mise à jour

est un résumé multi-document obtenu en condensant un ensemble de documents. Ce résumé suppose que le lecteur a déjà lu ou pris connaissances des documents sources.

En résumé automatique, on distingue deux grandes approches : le résumé par extraction et le résumé par abstraction. Le résumé extractif est le plus répandu, il consiste à extraire les phrases les plus importantes du texte pour former un résumé. Le résumé abstraitif consiste à extraire du document source les informations ou notions pertinentes afin de générer automatiquement un nouveau texte. Dans notre revue des méthodes du résumé automatique, nous nous sommes davantage intéressés au résumé extractif mono-document. Le résumé mono-document est un résumé produit à partir d'un document.

3.2 Les approches en résumé automatique

Deux approches sont distinguées en résumé automatique : les méthodes par extraction et les méthodes par abstraction.

3.2.1 Les méthodes par extraction

L'étape la plus cruciale est celle de la sélection des phrases. Ces dernières doivent être pertinentes et non redondantes. Les approches extractives créent un résumé en sélectionnant les phrases les plus importantes du document. Nous présentons maintenant les méthodes les plus utilisées en résumé par extraction :

- Méthodes fondées sur la fréquence des mots : Les travaux de Luhn (1958) ont exploré l'usage de la fréquence lexicale comme indicateur premier de l'importance de la phrase. Comme l'a fait remarquer Luhn (1958), certains mots les plus fréquents dans le document peuvent ne pas être indicatifs de ce qui est important car ils apparaissent souvent dans d'autres documents. L'approche du TF*IDF (Term Frequency * Inverse Document Frequency) a été proposée pour pallier au problème. Radev *et al.* (2004) proposent d'utiliser le Tf*IDF comme indice de pouvoir informatif des unités lexicales dans un système qui combine également une méthode permettant de sélectionner la phrase la moins semblable aux phrases déjà sélectionnées. Les phrases extraites sont celles qui possèdent les mots les plus informatifs du contenu du texte. Edmundson (1969) améliore l'approche de Luhn en définissant de nouvelles caractéristiques pouvant être utilisées pour indiquer l'importance d'une phrase. Ces caractéristiques comprennent la position de la phrase, la présence de mots-clés et de mots du titre, ainsi que d'entités nommées.
- Méthodes fondées sur les graphes : Elles permettent de modéliser le document selon une matrice d'adjacence où chaque nœud représente une phrase et le poids des arêtes entre chaque paire de phrases constitue leur similarité. Une procédure itérative est ensuite utilisée jusqu'à ce que les scores pour les arêtes convergent (Mnasri, 2015; Torres-Moreno *et al.*, 2015). Les travaux de (Radev *et al.*, 2004) ont montré que ces méthodes sont efficaces pour extraire les phrases les plus importantes du texte.
- Méthodes par apprentissage : Il s'agit essentiellement de méthodes supervisées qui visent à apprendre un modèle de choix de phrases à incorporer ou non au résumé. Un classifieur est entraîné à caractériser les phrases qui doivent apparaître dans le résumé et celles qui ne doivent pas l'être. Comme dans le cadre d'une tâche de classification, le classifieur permet de prédire les phrases devant être extraites pour composer le résumé. Ce problème a été soulevé

par Kupiec *et al.* (1995). Leurs travaux démontrent que l'apprentissage automatique combiné aux critères introduits par Luhn et Edmundson donne de bons résultats. Dans le cadre de leurs expérimentations, ils utilisent un classifieur Naïf Bayésien. D'autres ont également intégré les méthodes d'apprentissage automatique dans la tâche de résumé comme Neto *et al.* (2002) qui utilisent également un classifieur bayésien.

3.2.2 Les méthodes par abstraction

La construction d'un résumé par abstraction nécessite trois étapes : la représentation des connaissances, la sélection de ces connaissances et la génération du résumé à partir de ces connaissances (Bossard, 2010). Les méthodes abstractives apparaissent à la fin des années 1970. Le but de ces méthodes est de produire des résumés similaires à ceux que produiraient des humains. On distingue deux étapes en résumé abstractif : la compression du texte et la génération (Mnasri, 2015). La première analyse sémantiquement le contenu du texte soit par extraction, soit par regroupement de phrases. Trois méthodes peuvent être utilisées en résumé abstractif (Salima *et al.*, 2018) : les méthodes qui modélisent les informations contenues dans les textes en ontologies, patrons ou graphes, les méthodes utilisant la représentation sémantique des documents et les méthodes des réseaux de neurones.

3.3 Évaluation automatique du résumé

La tâche d'évaluation des résumés reste très difficile. Il n'existe pas vraiment de résumé de référence idéal pour un document donné compte tenu de la nature subjective de la tâche. Le développement des travaux en résumé automatique ont quand même permis la création de campagnes d'évaluations annuelles (TAC, TREC) afin de tester les performances des systèmes de résumé. Dans cette section, nous présentons quelques métriques d'évaluation des résumés.

3.3.1 Métriques d'évaluations du résumé automatique

- Précision, Rappel et F-mesure : si on considère la tâche de résumé automatique comme un problème de classification binaire des phrases à inclure ou à ne pas inclure dans le résumé final, des métriques d'évaluation comme la Précision, le Rappel et la F-mesure peuvent être utilisées. La Précision correspond au rapport du nombre d'énoncés sélectionnés par le système et présents dans le résumé de référence. Le Rappel est défini par le nombre d'énoncés pertinents retrouvés au regard du nombre d'énoncés pertinents que possède le résumé de référence. La F-mesure est la moyenne harmonique entre la précision et le rappel.
- ROUGE (Lin, 2004) est une métrique qui compare un résumé produit automatiquement à un résumé de référence ou à un ensemble de références (produit par l'homme). Il en existe plusieurs variantes : ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-SU.
- METEOR (Banerjee & Lavie, 2005) est une métrique pour l'évaluation des systèmes de traduction automatique. Elle peut également être utilisée pour la tâche du résumé automatique. Elle utilise la moyenne harmonique de la précision et du rappel des uni-grammes de la phrase. Elle a été proposée pour résoudre certaines faiblesses de la métrique BLEU (Papineni *et al.*, 2002), et également pour produire une bonne corrélation avec le jugement humain au niveau de la phrase ou du segment.

- PYRAMID (Nenkova & Passonneau, 2004) : l'idée derrière cette méthode est que la pertinence d'une unité d'information sémantique nommée SCU (Summary Content Unit) peut être déterminée par le nombre de résumés de référence qui l'incluent. Une SCU est une unité sémantiquement atomique représentant un fait unique, mais n'étant pas liée à sa réalisation lexicale ; deux paraphrases du même fait représentent la même SCU bien qu'elle soit exprimée différemment. Une SCU se voit attribuer un score proportionnel au nombre de résumés de référence qui la contiennent. Le score PYRAMID pour un résumé est calculé en prenant une moyenne normalisée des scores SCUs qu'il contient.

3.4 Travaux récents en résumé automatique appliqués aux corpus oraux transcrits

Historiquement, la tâche du résumé automatique a été appliquée à des données textuelles (articles journalistiques, scientifiques, bandes d'annonce de film, etc.) mais certains travaux se sont intéressés au résumé de conversations parlées transcrites.

Ganesh & Dingliwal (2019) proposent une nouvelle méthode pour le résumé par abstraction de conversations fondée sur les relations et la modélisation discursive. Le modèle proposé comprend l'étiquetage séquentiel des balises de prononciation, le réordonnement de la conversation pour modéliser les relations discursives, et un modèle de couverture pour le résumé abstraitif (*Pointer-Generator, Coverage based model*).

Dans leurs travaux axés sur la génération automatique de comptes-rendus, Alizadeh *et al.* (2018) utilisent la mesure de $TF*IDF$ et une approche de *topic modeling* s'appuyant sur l'algèbre linéaire (NMF⁷) pour la tâche d'extraction d'informations dans des retranscriptions de réunions. Le corpus utilisé est le corpus AMI (McCowan *et al.*, 2005).

Trione (2017) s'intéresse à la tâche du résumé automatique appliquée à des conversations issues des centres d'appels. L'objectif est de produire un résumé contenant les informations essentielles à la compréhension de l'appel. Il propose une méthode de résumé abstraitif par remplissage de patrons. Un ensemble de patrons de phrases est généré automatiquement à partir d'un corpus d'entraînement constitué de paires de transcriptions de conversations/résumés. Ensuite, les patrons sont sélectionnés en fonction des variables qui ont été détectées dans le document traité. Le résumé est généré par remplissage de ces patrons. Les conversations utilisées sont issues du corpus RATP-DECODA (Bechet *et al.*, 2012). Le corpus est composé de plus de 1500 conversations. Les versions audio et textuelle (transcriptions manuelles et automatiques) sont également disponibles. Les résultats montrent clairement que les transcriptions manuelles obtiennent de meilleurs résultats que les transcriptions automatiques. En outre, les résultats détaillés confirment que la méthode proposée par (Trione, 2017) produit de meilleurs résultats que les approches extractives (méthode MMR⁸) et les méthodes fondées sur l'utilisation de patrons figés (patrons écrits manuellement).

Dans leurs travaux, Wang & Liu (2015) se sont intéressés au résumé d'opinions appliqué à des conversations spontanées de locuteurs. Ils explorent deux méthodes. La première est une méthode non-supervisée fondée sur les graphes et la deuxième est une méthode supervisée : ils choisissent d'utiliser un classifieur SVM. La méthode fondée sur les graphes qu'ils utilisent incorpore la similarité cosinus entre deux phrases ($sim(s, v)$), la pertinence du sujet (REL), et le score de sentiment de

7. Non-negative Matrix Factorization

8. Maximal Marginal Relevance

la phrase (s : probabilité qu'une phrase contienne une opinion). En plus de ces trois scores, ils introduisent de nouvelles relations entre phrases pour modéliser la structure du dialogue. Pour la construction du classifieur SVM, ils utilisent quatre descripteurs : les trois premiers sont ceux utilisés pour l'approche fondée sur les graphes et le dernier prend en compte la longueur de la phrase. Ils décident également d'évaluer l'apport d'un module de résolution de coréférence et d'anaphore pronominale dans la production du résumé en introduisant trois nouveaux descripteurs : le contexte de référence, la pertinence de la référence (*REL*) et la chaîne de coréférence. La pertinence (*REL*) correspond au nombre de syntagmes nominaux qui renvoient à la même entité. Ils y ajoutent également un module de post-traitement qui remplace les pronoms par les unités lexicales auxquelles ils se réfèrent après sélection des phrases à introduire au résumé. Ils comparent ce système qu'ils nomment *SYS_PRON* à deux autres systèmes :

- une première baseline qui intègre les méthodes supervisées décrites ci-dessus.
- une deuxième baseline similaire à la première et intégrant le module de post-traitement.

Étant donné que les résumés extractifs de référence sont extraits des conversations originales contenant des pronoms, Wang & Liu (2015) proposent de les comparer avec les résumés abstractifs de référence. Les résultats montrent que le système *SYS_PRON* obtient les meilleurs résultats par rapport aux deux baselines démontrant ainsi que la prise en compte de la coréférence et de l'anaphore améliore les résumés produits automatiquement.

Zhu & Penn (2006) proposent un système de résumé automatique par extraction reposant sur une variété de descripteurs : le score MMR, les caractéristiques lexicales, structurelles, prosodiques et la présence de disfluences (les faux-départs, les phrases répétées ou recommencées, les syllabes répétées, les tics de langage, tels que les grommellements, propos non lexicaux tels que « euh », « hum », « hem »). Pour obtenir un module de sélection d'énoncés qui peut être entraîné et qui peut utiliser différents descripteurs, ils opèrent une classification binaire standard. Deux classifieurs sont choisis : le SVM et LR (Régression logistique). Ils utilisent à la fois les transcriptions automatiques et manuelles du corpus *Switchboard* (Godfrey *et al.*, 1992).

Furui (2005) propose une méthode de résumé en deux étapes consistant en une extraction de phrases importantes et une compression des phrases à partir de mots. Après avoir enlevé tous les *fillers*⁹ en se basant sur les résultats de la transcription automatique, un ensemble de phrases relativement importantes est extrait et une compression de phrases est appliquée à l'ensemble des phrases extraites. Le ratio d'extraction et de compression des phrases est contrôlé en fonction d'un ratio de compression initialement déterminé par l'utilisateur. Les unités de phrases et de mots sont extraites des résultats de la reconnaissance vocale et concaténées pour produire des résumés afin de maximiser la somme pondérée de la probabilité linguistique (*linguistic likelihood*), de la quantité d'information, de la mesure de confiance et de la probabilité grammaticale (*grammatical likelihood*) des unités concaténées. La méthode proposée a été appliquée pour résumer les journaux télévisés. La méthode a également été évaluée au moyen de mesures objectives et subjectives. Ces évaluations confirment que la méthode proposée est efficace pour la production de résumé.

En conclusion, les travaux en résumé automatique appliqués aux corpus oraux transcrits se sont focalisés sur les données de conversations téléphoniques (Ganesh & Dingliwal, 2019), de journaux télévisés (Furui, 2005) ou de réunions (Alizadeh *et al.*, 2018). Les premiers systèmes de résumé automatique de conversations exploitaient les méthodes extractives. Néanmoins, les méthodes abstractives énoncées ci-dessus fournissent elles aussi des résultats satisfaisants. Les transcriptions manuelles semblent se rapprocher des textes écrits, ce qui expliquerait la performance élevée des systèmes de

9. tic de langage

résumé automatique extractif alors que la présence d'erreurs liées à la reconnaissance vocale, de disfluences, d'hésitations, de répétitions, et de bruit dans les transcriptions automatiques sont autant d'éléments qui réduisent la précision des systèmes.

4 Conclusions et perspectives

Nous avons présenté deux domaines de recherche appliqués aux corpus oraux transcrits : l'analyse de sentiments et le résumé automatique. S'agissant de ces deux domaines de recherche, les travaux présentés ci-dessus se sont intéressés à divers types de données orales transcrites. À la différence des textes écrits, les conversations sont des échanges spontanés entre personnes. Ces échanges peuvent être altérés par la spontanéité des locuteurs et contenir du bruit. Nous étions particulièrement intéressés de savoir comment les auteurs traitaient les problèmes de l'oral pour ce type de données, les méthodes qu'ils proposaient ou mettaient en avant, et les traitements qu'ils préconisaient.

- *S'agissant du traitement des problèmes de l'oral dans les corpus*, il pose spécifiquement problème pour les transcriptions automatiques car elles sont plus sujettes à des erreurs de transcription, à la présence d'hésitations et de répétitions que les transcriptions manuelles qui peuvent être révisées ;
- *Au niveau du choix des descripteurs pour la construction des classifieurs*, d'autres caractéristiques telles que les informations contextuelles ou méta-données, les pronoms, les relations entre phrases de locuteurs différents ont été mises avant. Les résultats détaillés des articles ont montré qu'ils étaient cruciaux dans l'amélioration des systèmes de classification et d'extraction de phrases ;
- *Au niveau du choix des méthodes*, les méthodes supervisées sont les plus utilisées pour l'analyse de sentiments. Les méthodes extractives pour le résumé automatique sont privilégiées même si de nombreuses approches sur la compression automatique de phrases ont été mises en avant.

Au regard de ces deux états de l'art, nous estimons que nos travaux se différencient d'eux au niveau du matériau à analyser. Les données dont nous disposons sont des transcriptions manuelles d'entretiens semi-directifs non-révisées. Ce type de données est très peu étudié en TAL mais très présent dans le domaine des sciences humaines et sociales. Ces transcriptions se veulent plus fidèles à l'audio et intègrent donc les disfluences qui posent généralement des difficultés pour les systèmes de reconnaissance vocale. Il serait intéressant de se pencher sur des travaux qui se seraient intéressés à l'utilisation des disfluences dans la détection d'opinions ou le résumé automatique comme Zhu & Penn (2006).

Nous relevons aussi que très peu de travaux se sont intéressés au résumé d'opinions d'entretiens semi-directifs. Bien que les travaux de Wang & Liu (2015) ont porté sur le résumé d'opinion dans des conversations et ont permis de valoriser l'apport de l'analyse anaphorique dans la tâche du résumé, elles concernaient des conversations parlées issues de centre d'appels. Il serait également intéressant de pousser nos recherches sur les modules de résolution de coréférence et d'anaphores. Par ailleurs, nous souhaitons également nous tourner vers les travaux en résumé orienté et/ou guidé par la tâche. Les entretiens semi-directifs impliquent un échange entre deux personnes (la personne interrogée et la personne qui interroge). Cela permettrait d'évaluer la capacité de ces systèmes à condenser et à extraire les connaissances pertinentes contenues dans ce type de données.

Références

- ABDAOUI A., AZÉ J., BRINGAY S. & PONCELET P. (2017). Feel : a french expanded emotion lexicon. *Language Resources and Evaluation*, **51**(3), 833–855.
- ABDULLA N. A., AHMED N. A., SHEHAB M. A. & AL-AYYOUB M. (2013). Arabic sentiment analysis : Lexicon-based and corpus-based. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*, p. 1–6 : IEEE.
- ABERCROMBIE G. & BATISTA-NAVARRO R. (2018). 'aye'or'no' ? speech-level sentiment analysis of hansard uk parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- ALI C. B., MULKI H. & HADDAD H. (2018). Impact du prétraitement linguistique sur l'analyse des sentiments du dialecte tunisien. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, p. 383.
- ALIZADEH P., CELLIER P., CHARNOIS T., CRÉMILLEUX B. & ZIMMERMANN A. (2018). Étude expérimentale d'extraction d'information dans des retranscriptions de réunions. In *Traitement automatique du langage naturel (TALN)*.
- AUGUSTYN M., HAMOU S. B., BLOQUET G., GOOSSENS V., LOISEAU M. & RINCK F. (2006). Lexique des affects : constitution de ressources pédagogiques numériques. In *Colloque International des étudiants-chercheurs en didactique des langues et linguistique*.
- BANERJEE S. & LAVIE A. (2005). Meteor : An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, p. 65–72.
- BARHOUMI A., CAMELIN N. & ESTÈVE Y. (2018). Des représentations continues de mots pour l'analyse d'opinions en arabe : une étude qualitative. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, p. 215.
- BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). Decoda : a call-centre human-human spoken conversation corpus. In *LREC*, p. 1343–1347.
- BOSSARD A. (2010). *Contribution au résumé automatique multi-documents*. PhD thesis, Université Paris-Nord-Paris XIII.
- CAILLIAU F. & CAVET A. (2010). Analyse des sentiments et transcription automatique : modélisation du déroulement de conversations téléphoniques. *Revue Traitement Automatique des Langues*, **51**(3), 131–154.
- CAMELIN N., DAMNATI G., BÉCHET F. & DE MORI R. (2006). Détection automatique d'opinions dans des corpus de messages oraux. *Journées d'Etude sur la Parole (JEP'06), Dinard*.
- CHARAUDEAU P. (2001). *Grammaire du sens et de l'expression*. Hachette.
- CLICHE M. (2017). Bb_twtr at semeval-2017 task 4 : Twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv :1704.06125*.
- COLLINS E., AUGENSTEIN I. & RIEDEL S. (2017). A supervised approach to extractive summarization of scientific papers. *arXiv preprint arXiv :1706.03946*.
- DING X., LIU B. & YU P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, p. 231–240 : ACM.

- EDMUNDSON H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, **16**(2), 264–270.
- FURUI S. (2005). Spontaneous speech recognition and summarization. In *The Second Baltic Conference on Human Language Technologies*, p. 39–50.
- GAILLAT T., SOUSA A., ZARROUK M. & DAVIS B. (2018). Finsentia : Sentiment analysis in english financial microblogs. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, p. 271.
- GALA N. & BRUN C. (2012). Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions. In *19ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, volume 2, p. 495–502.
- GANESH P. & DINGLIWAL S. (2019). Abstractive summarization of spoken and written conversation. *arXiv preprint arXiv :1902.01615*.
- GODFREY J. J., HOLLIMAN E. C. & MCDANIEL J. (1992). Switchboard : Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, p. 517–520 : IEEE.
- GOYAL A. & DAUMÉ III H. (2011). Generating semantic orientation lexicon using large data and thesaurus. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, p. 37–43 : Association for Computational Linguistics.
- HEDAR A. R. & DOSS M. (2013). Mining social networks arabic slang comments. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*.
- KOOLI N. & PIGNEUL E. (2018). Analyse de sentiments à base d'aspects par combinaison de réseaux profonds : application à des avis en français. In *conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 1, p. 303–310.
- KUPIEC J., PEDERSEN J. & CHEN F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 68–73 : ACM.
- LAFOURCADE M., LE BRUN N. & JOUBERT A. (2015). Vous aimez?... ou pas ? likeit, un jeu pour construire une ressource lexicale de polarité. In *TALN : Traitement Automatique des Langues Naturelles*.
- LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- LIU B. & ZHANG L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, p. 415–463. Springer.
- LUHN H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, **2**(2), 159–165.
- MAUREL S., CURTONI P. & DINI L. (2008). A hybrid method for sentiment analysis. In *INFORSID*.
- MCCOWAN I., CARLETTA J., KRAAIJ W., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAIKOS V. *et al.* (2005). The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, p. 100.
- MNASRI M. (2015). Résumé automatique multi-document dynamique : État de l'art. *Actes des 17e Rencontres des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Association pour le Traitement Automatique des Langues, Caen*, p. 38–49.

- MOORE A. & RAYSON P. (2017). Lancaster a at semeval-2017 task 5 : Evaluation metrics matter : predicting sentiment from financial news headlines. *arXiv preprint arXiv :1705.00571*.
- MULKI H., HADDAD H. & GRIDACH M. (2017). Polarity analysis of non figurative tweets : Tw-star participation on deft 2017. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p.92.
- NENKOVA A., MCKEOWN K. *et al.* (2011). Automatic summarization. *Foundations and Trends® in Information Retrieval*, **5**(2–3), 104–141.
- NENKOVA A. & PASSONNEAU R. (2004). Evaluating content selection in summarization : The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics : Hlt-naacl 2004*.
- NETO J. L., FREITAS A. A. & KAESTNER C. A. (2002). Automatic text summarization using a machine learning approach. In *Brazilian Symposium on Artificial Intelligence*, p. 205–215 : Springer.
- NZALI M. D. T., ABDAOUI A., AZÉ J., BRINGAY S., LAVERGNE C., MOLLEVI C. & PONCELET P. (2017). Frenchsenticlass : un système automatisé pour la classification de sentiments en français. In *Actes de l'atelier DEFT de la conférence TALN 2017*.
- PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.
- PÉREZ-ROSAS V. & MIHALCEA R. (2013). Sentiment analysis of online spoken reviews. In *INTERSPEECH*, p. 862–866.
- POLAND B. D. (1995). Transcription quality as an aspect of rigor in qualitative research. *Qualitative inquiry*, **1**(3), 290–310.
- PONTIKI M., GALANIS D., PAPAGEORGIOU H., ANDROUTSOPOULOS I., MANANDHAR S., MOHAMMAD A.-S., AL-AYYOUB M., ZHAO Y., QIN B., DE CLERCQ O. *et al.* (2016). Semeval-2016 task 5 : Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, p. 19–30.
- RADEV D. R., JING H., STYŚ M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, **40**(6), 919–938.
- SALIMA L., SAID O. E. A. & BERNARD E. (2018). Résumé automatique guidé de textes : Etat de l'art et perspectives. *Actes des 20e Rencontres des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Association pour le Traitement Automatique des Langues, Rennes*.
- THOMAS M., PANG B. & LEE L. (2006). Get out the vote : Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, p. 327–335 : Association for Computational Linguistics.
- TORRES-MORENO J.-M., RAMIREZ J. & DA CUNHA I. (2015). Un résumeur à base de graphes, indépendant de la langue. *arXiv preprint arXiv :1501.01243*.
- TRIONE J. (2017). *Méthodes par abstraction et extraction pour le résumé de conversations orales issues de centres d'appels téléphoniques*. PhD thesis, Aix-Marseille.
- WANG D. & LIU Y. (2015). Opinion summarization on spontaneous conversations. *Computer Speech & Language*, **34**(1), 61–82.
- WIEBE J., WILSON T. & BELL M. (2001). Identifying collocations for recognizing opinions. In *Proceedings of the ACL-01 Workshop on Collocation : Computational Extraction, Analysis, and Exploitation*, p. 24–31.

ZHU X. & PENN G. (2006). Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.