Résumé automatique de textes d'opinion

Aurélien Bossard* — **Michel Généreux**** — **Thierry Poibeau*****

** Laboratoire d'Informatique de Paris-Nord (UMR 7030, CNRS et U. Paris 13)*
99, av. J.-B. Clément – 93430 Villetaneuse
aurelien.bossard@lipn.univ-paris13.fr

*** Centro de Linguistica da Universidade de Lisboa*
Av. Prof. Gama Pinto, 2 – 1649-003 Lisboa – Portugal
genereux@clul.ul.pt

**** Laboratoire LaTTiCe (UMR 8094, CNRS, ÉNS et U. Paris 3)*
1, rue Maurice Arnoux – 92120 Montrouge
thierry.poibeau@ens.fr

RÉSUMÉ. *Nous présentons dans cet article un système de résumé automatique tourné vers l'analyse de blogs, où sont exprimées à la fois des informations factuelles et des prises de position sur les faits considérés. Notre système de résumé est fondé sur une approche nouvelle qui mêle analyse de la redondance et repérage des informations nouvelles dans les textes ; ce système générique est en outre enrichi d'un module de calcul de la polarité de l'opinion véhiculée afin de traiter de façon appropriée la subjectivité qui est le propre des billets de blogs. Le système est évalué sur l'anglais, à travers la participation à la campagne d'évaluation internationale TAC (*Text Analysis Conference*) où notre système a obtenu des performances satisfaisantes.*

ABSTRACT. *In this paper, we present a summarization system that is specifically designed to process blog posts, where factual information is mixed with opinions on the discussed facts. Our approach combines redundancy analysis with new information tracking and is enriched by a module that computes the polarity of textual fragments in order to summarize blog posts more efficiently. The system is evaluated against English data, especially through the participation in TAC (Text Analysis Conference), an international evaluation framework for automatic summarization, in which our system obtained interesting results.*

MOTS-CLÉS : *résumé automatique, analyse de blogs, analyse de l'opinion, redondance, subjectivité, évaluation de résumés automatiques, évaluation de l'analyse de l'opinion.*

KEYWORDS: *automatic summarization, analysis of blogs, opinion mining, redundancy, subjectivity, automatic summarization evaluation, opinion analysis evaluation.*

1. Introduction

Le résumé automatique a connu un fort renouveau ces dernières années. Si les recherches menées dans ce domaine s'inscrivent dans une tradition longue de plus de 50 ans (Luhn, 1958), elles ont fortement évolué récemment : l'apparition de gros corpus parfois hétérogènes et la généralisation des techniques d'analyse de surface ont à la fois renouvelé les besoins et les approches. Plus récemment encore, avec l'avènement de médias plus interactifs (le fameux Web 2.0), la nécessité de repérer les citations, les jugements et les opinions s'est avérée de plus en plus cruciale. En effet, de telles données, aujourd'hui facilement accessibles, peuvent aider à comprendre les besoins et les attentes de toute une population, et également à analyser les opinions concernant des produits, des personnalités ou même des propositions politiques.

Le but n'est plus seulement de produire une synthèse de l'information contenue dans les textes, il faut en outre dégager des tendances, identifier les opinions exprimées et si possible en faire la synthèse. Cet article décrit des recherches menées dans ce cadre : nous avons développé un système visant à faire une synthèse automatique des opinions exprimées sur Internet sur un sujet donné, en particulier dans des blogs. Le système repose sur une synthèse entre des techniques de production de résumés par extraction de passages pertinents et l'analyse de la polarité des opinions exprimées dans ces textes.

Pour évaluer notre système, nous avons participé à la campagne TAC 2008[1] (*Text Analysis conference*), une campagne d'évaluation internationale organisée par le NIST (*National Institute of Standards and Technology*) et tournée vers les systèmes de questions-réponses (QR) et de résumé automatique. L'évaluation proposée dans ce cadre mêlait résumé factuel et analyse d'opinion, à partir des sorties de systèmes de recherche d'information[2]. Le fonds documentaire était constitué de blogs et l'enjeu était de produire des synthèses cohérentes à partir de questions en langage naturel – en général, un résumé correspond à plusieurs questions liées (appelées *squishy list*) sur un thème donné (appelé *target*).

Pour prendre un exemple, un des thèmes proposés visait la personnalité de l'année désignée par le magazine *Time* pour 2005 (« *Time Magazine 2005 Person of the Year* »). Les questions liées étaient les suivantes : « *Why did readers support Time's inclusion of Bono for Person of the Year ?* », « *Why did readers not support the inclusion of Bill Gates as Person of the Year ?* », « *Why did readers not support the inclusion of Melinda Gates as Person of the Year ?* ». On voit qu'il s'agit de questions en « pourquoi » (*why*) : contrairement aux questions factuelles (questions dites

1. La campagne sur le résumé d'opinion n'a pas eu lieu en 2009.

2. Il était aussi possible de participer en utilisant les sorties de systèmes de questions-réponses ; nous avons participé aux deux sous-tâches mais le système présenté ici utilise seulement les sorties d'un système de recherche d'information (c'est-à-dire des ensembles de documents *a priori* pertinents par rapport à une question donnée) et non ceux d'un système de QR (on ne peut donc ici utiliser les *snippets*, ces séquences de caractères extraites du fonds documentaire par les systèmes de QR et censées apporter des éléments de réponse aux questions posées).

factoïdes, où la réponse est généralement une entité nommée), il n'est pas possible de répondre de façon simple à ces enchaînements de questions en pourquoi. Les systèmes de questions-réponses traditionnels, qui produisent des fragments en guise de réponse (*snippets*) sont insuffisants dans ce cadre, dans la mesure où ils ne permettent pas de « contextualiser » correctement la réponse, c'est-à-dire de produire un tout cohérent rendant compte des opinions exprimées. La production de résumés à partir d'une extraction de phrases donnant une idée des informations essentielles contenues dans le fonds documentaire et formant autant que possible un tout cohérent, semble une voie plus prometteuse.

Dans ce cadre, le résumé doit être tourné vers les séquences exprimant une opinion. Un tel système se différencie d'un système de résumé classique par l'intégration d'un module d'analyse d'opinion, ou plus exactement, de la polarité des phrases (phrases exprimant un opinion positive, négative ou neutre). Mais cette analyse de la polarité ne doit pas cacher l'essentiel : un bon résumé doit avant tout refléter les arguments exprimés, qu'ils soient positifs ou négatifs. Nous détaillerons donc l'architecture de notre système appelé CBSEAS (Bossard *et al.*, 2008), qui permet de rendre compte de la diversité informationnelle exprimée dans les textes à résumer. Cette diversité permet de reconnaître les différents arguments exprimés. Le module d'analyse d'opinion agit alors comme un filtre sur le système de résumé, en ne gardant que les phrases ayant la même polarité que la requête[3].

L'article essaie de montrer comment les différents aspects du problème sont gérés et intégrés dans notre système. Nous présentons également l'évaluation proposée dans le cadre de la campagne TAC. Si l'évaluation du résumé multidocument est un problème ouvert, on y ajoute une difficulté supplémentaire quand se superpose la question de l'analyse d'opinion. Nous présentons donc à la fois les résultats obtenus et une discussion autour de ces résultats, dans la mesure où les chiffres officiels fournis doivent être discutés en détail, vu le caractère exploratoire de ces recherches.

Le plan de l'article est comme suit. Après un état de l'art du domaine, nous présentons le système que nous avons développé pour le résumé automatique de textes véhiculant une opinion ; ses différents aspects sont détaillés, notamment les indices de surface utilisés, les mesures de calcul de proximité sémantique et les procédures de choix des phrases extraites et leur ordonnancement. Nous présentons ensuite la tâche « résumé d'opinion » de la campagne TAC 2008, les résultats que nous avons obtenus et leurs limites.

3. On exclura donc du résumé les phrases exprimant une opinion contraire à celle souhaitée dans la question. Par exemple, si la question porte sur les raisons de soutenir la candidature de R. Giuliani à la mairie de New York, on exclura les opinions négatives sur R. Giuliani, celles-ci n'étant pas pertinentes dans le contexte visé.

2. État de l'art

Nous présentons ici un aperçu des techniques de résumé multidocument et des techniques d'analyse d'opinion. Vu le foisonnement de recherche dans ces deux domaines, nous limitons volontairement notre aperçu à des références directement utiles par rapport au système présenté par la suite. En ce qui concerne l'analyse d'opinion, nous citons des travaux visant à identifier la polarité de séquences textuelles et l'extraction d'opinion, dans la mesure où c'est essentiellement ce type de technique qui est mis en œuvre dans le contexte qui nous occupe.

2.1. *Du résumé multidocument au résumé de textes d'opinion*

La production automatique de résumé est une tâche ancienne qui remonte aux débuts du traitement automatique des langues par ordinateur. Dès les années 1950, les ordinateurs semblaient pouvoir répondre aux besoins en matière d'information dans deux secteurs complémentaires, d'une part en pouvant automatiser la production de traduction (Hutchins et Somers, 1992), d'autre part en aidant l'analyse documentaire (Spärck Jones et Willett, 1997).

Le résumé est donc d'emblée apparu comme un domaine de choix pour l'analyse automatique. Dès les premières tentatives, les chercheurs ont essayé de définir des approches simples visant à extraire des phrases pertinentes (ou des segments de phrases) qui, organisées ensemble, devaient malgré tout donner un résultat compréhensible et facilement lisible par un humain. On a alors obtenu ce qui est appelé en anglais des *extracts*, c'est-à-dire des résumés composés de phrases extraites du document original (Luhn, 1958 ; Edmundson, 1969). Ces techniques reposent sur l'identification des éléments les plus porteurs de sens, après un travail d'analyse manuel.

Afin de rendre l'approche plus portable, les concepteurs de systèmes récents ont essayé de ne plus avoir à définir à la main les éléments pertinents mais de les déterminer automatiquement. Si l'on dispose d'un ensemble de textes représentatifs et de résumés préalablement élaborés manuellement, le système doit pouvoir déterminer quels sont les mots ou les expressions discriminantes pour la sélection de phrases, sans intervention extérieure. Diverses variantes de cette approche ont été proposées pour identifier les phrases les plus centrales (Radev *et al.*, 2001 ; Radev *et al.*, 2004 ; Boudin et Torres-Moreno, 2007) et pour éliminer la redondance (Carbonell et Goldstein, 1998 ; Boudin *et al.*, 2008).

La plupart des expériences faites jusqu'au début des années 2000 portaient sur des textes de presse, où les techniques génériques sont relativement efficaces. Mais on voit alors apparaître des besoins plus spécialisés, notamment en ce qui concerne l'analyse d'opinion. Cardie *et al.* (2003) proposent ainsi un schéma de système (non implémenté) intégrant un système de questions-réponses, un système de résumé et un module d'analyse d'opinion, pour répondre à des questions du type de celles que nous avons vues dans l'introduction.

Parmi les travaux pionniers, on peut citer Hu et Liu (2004) qui proposent un système offrant une visualisation des opinions exprimées sur un objet donné. Cette expérience est intéressante mais elle n'aborde pas la question de la sélection des phrases et de leur ordonnancement vu que la sortie du système est juste un graphique exprimant les tendances de l'opinion.

Des systèmes récents visent à produire des résumés textuels intégrant l'analyse d'opinion. Les informations contenues dans ces résumés doivent permettre de saisir à la fois les polarités d'opinion exprimées et leurs causes. La campagne d'évaluation organisée sur ce thème pour TAC 2008, que nous présentons en détail en section 5.1, cherchait ainsi à évaluer des systèmes de résumé textuel orienté vers une polarité d'opinion donnée, à propos d'un sujet précis (qu'il s'agisse d'un objet, d'une personne, ou encore d'un fait d'actualité). Dans les différents systèmes ayant participé, l'intégration de l'analyse d'opinion se fait majoritairement en filtrant, pour une requête donnée, les phrases véhiculant l'opinion demandée (Seki, 2008 ; Kim et Zhai, 2009). L'analyse de l'opinion peut aussi être directement intégrée au score affecté à chaque phrase pour évaluer sa pertinence par rapport à la tâche (Murray *et al.*, 2008). Le risque est alors de faire passer l'analyse de l'opinion au second plan et de produire des résumés trop focalisés sur la requête et véhiculant moins de subjectivité.

L'approche que nous proposons ici se démarque de l'existant par l'utilisation d'une technique novatrice de résumé automatique, qui vise à maximiser la diversité de l'information retenue. L'analyse d'opinion intervient alors comme un filtre. Ceci semble particulièrement pertinent dans notre cadre d'étude, puisqu'il est nécessaire de pouvoir extraire un maximum de faits supportant l'opinion exprimée (positive ou négative).

2.2. *Analyse de la polarité de séquences textuelles*

L'analyse de l'opinion véhiculée dans les textes est devenue un domaine de recherche très actif ces dernières années. On distingue trois sous-tâches principales. La première sous-tâche consiste à distinguer les textes subjectifs des textes objectifs (Bethard *et al.*, 2004) ; la deuxième s'attarde à classer les textes subjectifs en positifs ou négatifs (Turney, 2002) ; enfin, la troisième essaie de déterminer jusqu'à quel point les textes sont positifs ou négatifs (Wiebe *et al.*, 2001).

Plusieurs ressources ont été développées récemment autour de l'analyse d'opinion, ou, plus largement, de tout ce qui concerne les sentiments face à un événement ou une situation donnée. On pourra citer Wordnet-Affect (Strapparava et Valitutti, 2004) ou SentiWordnet (Esuli et Sebastiani, 2006 ; Baccianella *et al.*, 2010) pour l'anglais. La première ressource est plus large que la seconde, dans la mesure où elle couvre une large variété de sentiments, tandis que la seconde est davantage orientée vers l'analyse d'opinion. Pour le français, on pourra citer le lexique développé par Y. Mathieu, qui vise à répartir les termes simples exprimant un sentiment en 38 classes sémantiquement homogènes (Mathieu, 2005). Des méthodes semi-automatiques destinées à compléter les ressources manuelles ont été conçues plus récemment (Vernier et Mon-

ceaux, 2007). Une étape importante pour la création de résumés de textes subjectifs est l'identification des passages contenant des opinions (Kao et Chen, 2010) et Twitter fait déjà son apparition comme une source pertinente pour l'élaboration de corpus d'opinion (Pak et Paroubek, 2010). Ces derniers travaux offrent des perspectives d'amélioration importantes pour un système comme celui que nous présentons.

Afin d'avoir un système plus facilement adaptable, nous avons de notre côté préféré ne pas nous appuyer directement sur ce type de données mais prévoir un mécanisme d'apprentissage dynamique à partir de corpus. Il a été montré que ce type de démarche est efficace quand on ne vise qu'une analyse simple de la polarité des phrases (Turney, 2002).

Comme nous l'avons vu dès l'introduction, notre traitement des opinions pour l'aide à la production de résumés s'attache avant tout à repérer la distinction standard positif *versus* négatif. Il faut signaler les efforts récents pour réintroduire des approches plus linguistiques et discursives (prise en compte de la modalité, de l'énonciateur) dans ce domaine (Asher *et al.*, 2008). Ces recherches vont toutefois au-delà de ce qu'il est actuellement possible d'intégrer efficacement dans les systèmes de résumé, vu le nombre de paramètres à prendre en compte.

Il faut enfin noter l'impulsion donnée par des campagnes telles que TREC Blog Opinion Task depuis 2006 (Zhang *et al.*, 2007 ; Dey et Haque, 2008) : ces campagnes ont entraîné une forte opérationnalisation du domaine et l'intégration des techniques d'analyse d'opinion dans des systèmes ouverts.

3. Un système de résumé générique : CBSEAS

Les systèmes de production de résumés par extraction se fondent pratiquement tous sur le repérage des phrases supposées les plus importantes (« centrales ») dans les textes sources ; au sein de celles-ci, les phrases trop similaires (donc possiblement redondantes) sont progressivement éliminées, jusqu'à ce que le nombre de phrases restantes corresponde à la taille du résumé visé. L'identification des phrases centrales peut se faire par analyse du corpus (Radev *et al.*, 2001 ; Boudin et Torres-Moreno, 2007), ou par rapport à une requête quand le résumé doit répondre à une question utilisateur (Carbonell et Goldstein, 1998 ; Boudin *et al.*, 2008).

Contrairement à ces approches, nous présentons ici une technique qui vise à établir un modèle de représentation qui se fonde largement sur la notion de redondance. Il s'agit de rendre compte à la fois de la diversité et de la centralité des phrases, avant de sélectionner sur cette base celles à extraire.

3.1. *Un système fondé sur l'analyse de la redondance*

Générer automatiquement des résumés multidocuments ajoute une problématique supplémentaire à la génération de résumés monodocuments : l'élimination de la redon-

– As for George Clooney, it's a well known fact that he's a bad actor who gets by on **his good looks and charm**.
– He only became popular because the excellent writers on that still-successful show played to the sole strengths Clooney has as an actor : **his good looks and charm**.
– The 44-year-old star, known for **his easy charm** [...]
– "George Clooney : **Good looking**, politically savvy–even the nurses like him"
– "Dashing movie-star **good looks** and the Clooney-dreamy-quotient aside"

Source : corpus TREC Blogs06 (corpus ayant servi de support à la campagne TAC Opinion Summarization 2008)

Figure 1. *Rapport entre la centralité et la redondance : en gras, l'information centrale des différents extraits. On voit que cette information doit être contextualisée pour faire sens (les deux premiers extraits sont plutôt négatifs, les autres plutôt positifs).*

dance. L'information est en effet plus redondante si le système s'appuie sur plusieurs documents plutôt que sur un document unique. Le risque d'extraire plusieurs phrases véhiculant la même information est alors plus élevé.

Cette redondance apporte cependant des informations précieuses pour la réalisation du résumé. En effet, les informations les plus centrales ont de fortes chances d'être reprises dans plusieurs documents (*cf.* figure 1). Par conséquent, réussir à identifier les passages redondants permet non seulement l'élimination de la redondance du résumé final – un des critères de qualité des résumés –, mais également une meilleure sélection des phrases à extraire.

Notre système doit donc être fondé sur la détection de la redondance puis sur l'extraction des phrases les plus centrales au sein des classes de phrases redondantes (Erkan et Radev, 2004). Cela permet une élimination efficace de la redondance et la création de résumés riches du point de vue de la diversité informationnelle. Nous avons cependant évité les traitements trop lourds, comme une analyse syntaxique, qui risquerait de toute manière d'être peu performante sur les blogs (certains blogs sont en effet écrits dans une langue peu académique, avec peu ou pas de ponctuation). Nous en sommes donc restés à des traitements de surface facilement portables d'un domaine à l'autre.

3.2. *Architecture de CBSEAS*

CBSEAS a été pensé de manière à être aisément adaptable aux différentes tâches du résumé automatique, en reprenant plusieurs éléments de la proposition d'architecture générique présentée dans (Mani et Maybury, 1999). Nous distinguons quatre modules :

1) préparation des documents et des éventuelles requêtes utilisateur (*cf.* § 4.1) ;

2) construction d'un modèle pour représenter les phrases des documents et prendre en compte centralité et diversité (*cf.* section § 4.2). Ce modèle est fondé sur *i*) une analyse du contenu des phrases (*cf.* section § 4.2.1), *ii*) une fonction de similarité entre phrases sur la base de l'analyse faite précédemment (*cf.* section § 4.2.2) et *iii*) le regroupement de celles-ci en classes sémantiques (*cf.* section § 4.2.3) ;

3) extraction des phrases en s'appuyant sur la modélisation produite précédemment et en intégrant l'analyse de la polarité d'opinion (*cf.* § 4.3) ;

4) post-traitements, afin de gérer notamment l'ordre des phrases et la longueur du résumé final (*cf.* § 4.4).

Ce modèle nous permet d'obtenir une vue sur la redondance *via* les classes sémantiques (chaque classe représente un ensemble de phrases fortement redondantes). La similarité entre les phrases d'une même classe nous permet d'évaluer la centralité des phrases dans leur classe (plus une phrase est centrale, plus la probabilité est grande qu'elle véhicule l'information essentielle de cette classe ; c'est donc elle qui doit être retenue pour être insérée dans le résumé). Associer cette notion de centralité à d'autres paramètres utilisateur permet au système de sélectionner les phrases à la fois centrales et correspondant aux besoins de l'utilisateur (comme une requête) ou des demandes de résumés particuliers (comme des résumés d'opinion).

4. Application au résumé d'opinion : le système CBSEAS-Opinion

Nous détaillons dans cette section l'architecture précise de notre système et les spécificités de l'analyse d'opinion à partir de blogs. L'implémentation a permis de produire le système CBSEAS-Opinion.

4.1. *Nettoyage des documents*

Le contenu de blogs est souvent « pollué » par la publicité mais aussi par des messages complètement hors sujet. Afin de remédier à ces deux problèmes, nous avons choisi d'adopter une solution fondée sur le vocabulaire des phrases. Nous éliminons toute phrase dont le ratio :

nombre de mots fréquents / nombre total de mots

est inférieur à un seuil (0,35). Nous nous fondons sur les 100 mots les plus fréquents de l'anglais, qui constituent approximativement la moitié des textes écrits (Fry *et al.*, 2000). Le seuil de 0,35 a été fixé empiriquement en essayant différents seuils sur un ensemble de billets de blogs. Il permet une certaine souplesse quant à la sélection de

phrases qui n'emploient pas un vocabulaire courant, tout en éliminant celles qui sont écrites dans un anglais plus qu'approximatif[4].

Une seconde phase de présélection des phrases est alors appliquée, afin de gagner en efficacité lors des étapes ultérieures. Après avoir éliminé les mots vides des questions pour le sujet du résumé[5], le système ne garde que les phrases qui contiennent au moins un mot issu des questions posées par l'utilisateur. Nous faisons l'hypothèse que les autres phrases ne sont pas pertinentes : en effet, les questions contiennent toujours un objet central, qui est généralement une entité nommée. Ne garder que les phrases qui contiennent au moins un mot de la question permet de minimiser le risque de travailler sur des phrases dont la centralité vis-à-vis de la question n'est pas avérée.

4.2. *Modèle de représentation du contenu des phrases*

Les phrases restantes sont alors analysées afin de déterminer des classes d'équivalence, c'est-à-dire des ensembles de phrases véhiculant des informations similaires.

4.2.1. *Traitements linguistiques et analyse phrastique*

Les documents en entrée du système subissent les traitements préalables suivants.

– **Annotation morphosyntaxique :** les documents sont analysés morphosyntaxiquement, et les unités textuelles annotées par Treetagger[6] (Schmid, 1994). Ce traitement permet de différencier les différents types morphosyntaxiques dans les calculs.

– **Découpage des documents en phrases :** certains systèmes de résumé automatique font le choix de ne pas travailler sur les phrases, mais sur des structures plus petites. Ils travaillent alors à l'extraction de groupes de mots reliés syntaxiquement, en découpant les phrases en propositions (Marcu, 2000). D'autres systèmes ont choisi d'extraire des paragraphes entiers afin de garder une cohérence linguistique au sein du résumé. Le risque est alors plus élevé d'extraire des phrases non pertinentes puisque c'est le paragraphe dans son ensemble dont la pertinence est évaluée. Afin d'éviter cet écueil mais également de ne pas prendre le risque d'extraire des propositions mal découpées qui perdraient alors leur sens et leur grammaticalité, nous avons fait le choix d'extraire uniquement des phrases entières.

– **Calcul d'un score centroïde :** un score dit « centroïde » est attribué à chaque phrase. Nous calculons ce score d'après la méthode décrite dans (Radev *et al.*, 2001). Le centroïde est un groupe de mots composé des n mots les plus importants des jeux de documents à résumer. Ces mots sont affectés d'un score selon leur centralité ; ce score correspond à leur *tf.idf* ((Salton et Buckley, 1988), *cf.* figure 2) qui permet de mettre

4. Cette manière de filtrer a aussi l'avantage de ne pas éliminer les phrases courtes si elles sont bien formées.

5. Nous avons utilisé la liste des mots vides pour l'anglais issue de `http://www.textfixer.com/resources/common-english-words.txt`.

6. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

en évidence les termes les plus caractéristiques d'un sous-corpus donné au sein d'un corpus plus grand. Ici, chaque groupe de documents lié à une question particulière constitue un sous-corpus qui peut être opposé à l'ensemble des documents associés aux différentes questions. Finalement, le score centroïde d'une phrase est la somme des *tf.idf* communs à la phrase et au centroïde. Ce score est fonction de la centralité d'une phrase vis-à-vis du contenu global des documents à résumer.

Calcul du tf d'un terme t_i dans le document d_j :
$\text{tf}_{t_i,d_j} = \frac{n_{i,d_j}}{\sum_{k=0}^{n} n_{k,d_j}}$
où $n_{i,d}$ est le nombre d'occurrences du terme i dans le document d.

Calcul de l'idf d'un terme t_i au sein d'un corpus D de documents d_j :

$\text{idf}_{t_i} = \log \frac{|D|}{|d_j : t_i \in d_j|}$

Calcul du tf.idf du terme t_i pour un document d_j :
$\text{tf.idf}_{t_i,d_j} = tf_{t_i,d_j} \times idf_{t_i}$

Figure 2. *Calcul du* tf.idf

4.2.2. *Calcul des similarités entre phrases*

Nous faisons l'hypothèse que la similarité entre phrases doit tenir compte du type des documents que CBSEAS a à résumer. Par exemple, les caractéristiques qui détermineront si deux phrases sont similaires dans le cadre de l'analyse d'opinion diffèrent des caractéristiques qui détermineront la similarité de deux phrases issues d'un corpus boursier. Dans le premier cas, les mots exprimant un sentiment seront discriminants, tandis que dans l'autre cas ce seront les devises et montants, les verbes d'action et les entités nommées de type COMPAGNIE ou GROUPE.

Nous voulons rendre compte de ce fait au travers d'une mesure de similarité paramétrable, adaptable aux différentes tâches qui peuvent être proposées dans le large domaine du résumé automatique. Nous avons opté pour une mesure de comparaison d'ensembles du type Jaccard, pondérée par une valeur dépendant du type des termes comparés.

Le caractère discriminant des termes dans le cadre d'une mesure de similarité entre phrases ne dépend pas seulement de leur catégorie morphosyntaxique, mais également de leur importance relative à l'ensemble des documents. Certaines approches, comme celle de (Saggion, 2005), utilisent pour cela la pondération *idf (inverse document frequency)* calculée sur un corpus comme le *BNC (British National Corpus)*, assez large pour que les mesures *idf* soient valides. Cependant, une telle méthode, qui est viable pour la langue générale, devra être révisée pour chaque application à un domaine de

spécialité, ou à une langue différente. Pour cette raison, nous avons choisi de pondérer les termes par leur *tf.idf* (*cf.* figure 2) dans les documents à résumer, qui reflète bien leur importance vis-à-vis de ceux-ci[7].

$$sim(p1, p2) = \frac{\sum_{m \in p1 \cap p2} poids(t_m) \times tf.idf(m)}{\sum_{m \in p1 \cup p2} poids(t_m) \times tf.idf(m)}$$

Figure 3. *Indice de Jaccard utilisé pour la comparaison entre phrases*

La mesure de similarité qui résulte de ces deux pondérations est présentée en figure 3. Cette comparaison est fondée sur l'analyse des éléments pertinents m (mots sémantiquement pleins, termes, entités nommées) de $p1$ et/ou de $p2$ (les phrases à comparer). Chaque élément est en outre pondéré selon sa catégorie (t_m), ce qui permet de distinguer d'éventuels homonyme, voire des mots polysémiques.

4.2.3. *Classification des phrases en classes sémantiques*

Une fois la matrice de similarité établie, CBSEAS regroupe les phrases similaires. Cette étape s'effectue grâce à l'algorithme *fast global k-means* (Likas *et al.*, 2001) (*cf.* figure 4), une variante itérative de *k-means*. Cet algorithme de classification a l'avantage d'être itératif, et de pouvoir être facilement adapté à des tâches de résumé avec ajout de groupes de documents à la volée, comme les tâches de mises à jour des campagnes d'évaluation TAC (Bossard *et al.*, 2008). Il permet également de s'affranchir de la sélection des centres de classes préalablement à la classification, un défaut majeur de *k-means*. De plus, *fast global k-means* fonde les regroupements sur les similarités ou dissimilarités entre les éléments. Cela nous permet d'utiliser aisément différentes mesures de similarité. *k-means* (Forgy, 1965 ; MacQueen, 1967), ou k-moyennes, réalise une classification en k classes en minimisant la variance intra-classe. Cet algorithme comprend quatre étapes :

1) choisir aléatoirement k objets qui seront les centres de k classes ;

2) parcourir tous les objets, et les affecter ou les réaffecter à la classe qui minimise la distance entre l'objet et le centre de la classe ;

3) calculer les barycentres de chaque classe, qui deviennent les nouveaux centres ;

4) répéter les étapes 2 et 3 jusqu'à convergence. La convergence est atteinte lorsque les classes deviennent stables.

L'algorithme *fast global k-means* commence par créer une classe, dans laquelle sont placés tous les éléments (les phrases dans notre cas). À chaque itération, l'algorithme crée une nouvelle classe dont le centre sera l'élément le moins bien représenté

7. La mesure de similarité *cosinus* a également été testée, et s'est révélée être dans le même ordre de performance.

$i = 0$
k = nombre de clusters souhaité (avec k >= 2)
Pour tous les p_m dans P (l'ensemble des phrases à classer)
  $C_1 \leftarrow p_m$ (toutes les phrases p_m sont initialement incluses dans un même cluster C_1)
Faire
  Incrémenter i
  Ajouter un nouveau cluster C_i qui contient la phrase la plus éloignée du centre de son cluster actuel ; cet élément devient le centre de C_i
  Reclasser l'ensemble des p_m de P tel que la phrase p_m est classée dans le cluster C dont elle est le plus proche (autrement dit, le cluster C choisi sera celui pour lequel la distance entre le centre de C et p_m sera minimale)
  Recalculer l'ensemble des centres des i clusters (le centre d'un cluster est la phrase qui minimise la distance moyenne avec l'ensemble des autres phrases du cluster)
Tant que i <= k

Figure 4. *Algorithme* Fast global k-means *; la distance d'une phrase par rapport au centre d'un cluster est calculée en utilisant l'indice de Jaccard détaillé* supra.

par sa classe ; chaque élément est alors placé dans la classe dont il est le plus proche ; pour finir, les centres des classes sont recalculés.

Les regroupements étant établis d'après la matrice de similarité, elle-même uniquement établie en fonction du lexique des phrases, utiliser l'adjectif « sémantique » pour décrire ces regroupements pourrait paraître abusif. Cependant, étant donné que le sens d'un mot est fortement dépendant de son contexte, la probabilité est élevée pour que deux phrases qui partagent les mêmes mots partagent également le même sens (Gale *et al.*, 1992).

4.3. *Sélection des phrases*

L'étape suivante consiste à sélectionner les phrases les plus pertinentes. Le système CBSEAS a une procédure fondée sur la notion de centralité, qui est exposée ci-dessous. Pour la campagne TAC sur le résumé d'opinion, nous avons juste ajouté un module permettant de ne retenir *in fine* que les phrases ayant une polarité identique à la requête. L'adaptation du système à la tâche est donc légère et ne nécessite pas de modifier CBSEAS en profondeur.

4.3.1. *Analyse de la « centralité » des phrases*

Après avoir construit la représentation informatique des documents à résumer, le système CBSEAS extrait les phrases qu'il juge les plus pertinentes pour constituer un texte qui servira de base au résumé final. Afin d'obtenir un résumé qui maximise la

diversité informative, CBSEAS extrait une phrase par classe sémantique. La phrase sélectionnée doit maximiser un score issu des deux caractéristiques suivantes :

1) la centralité locale (la centralité d'une phrase dans sa classe sémantique) ;

2) la centralité globale (la centralité d'une phrase vis-à-vis d'une requête utilisateur, ou d'un *centroïde* de l'ensemble des documents).

Chaque classe est censée refléter un pan de la diversité informationnelle des documents à résumer. Extraire une phrase parmi chacune des classes permet d'obtenir un résumé qui maximise cette diversité, tout en maximisant également la centralité vis-à-vis de la requête et du centroïde.

La mesure de la centralité locale vise à déterminer si une phrase est représentative du contenu sémantique de sa classe, tout comme les mesures de centralité globale rendent compte de la représentativité du contenu global des documents. Nous partons du principe suivant : chaque phrase exprime une ou plusieurs « informations atomiques »[8]. L'ensemble des phrases d'une classe exprime ainsi un ensemble I d'informations atomiques. Ainsi, la phrase la plus centrale de cette classe sera celle qui exprime les informations les plus importantes de I. Nous travaillons d'après l'hypothèse que les informations les plus redondantes sont les plus importantes. Ainsi, la phrase qui contient les informations les plus répétées dans sa classe sera la plus centrale. La mesure de centralité que nous utilisons est avant tout fondée sur la comparaison d'éléments lexicaux et d'entités nommées.

La mesure de centralité locale permet de gérer le problème de la diversité en attribuant des scores élevés aux phrases les plus centrales dans leur classe. Il faut par ailleurs rendre compte de la centralité des phrases extraites vis-à-vis du contenu global des documents à résumer et vis-à-vis de la requête utilisateur éventuelle. Pour ce faire, nous utilisons les scores calculés lors de la préparation des documents (*cf.* § 4.2.1).

4.3.2. *Analyse de la polarité des phrases*

Nous avons tenté de donner à chaque phrase une polarité positive, négative ou neutre, en utilisant une approche supervisée qui s'appuie sur un jeu de documents étiquetés (Généreux, 2009). Deux classifieurs SVM (Joachims, 1997) ont été entraînés sur ce jeu de documents, un pour catégoriser les requêtes, et l'autre pour catégoriser les phrases issues des billets des blogs à résumer.

La détection de la polarité des requêtes est destinée à orienter au maximum le résumé vers les attentes de l'utilisateur. Si la question « *What did American voters admire about Rudy Giuliani ?* » est posée, il faudra détecter que l'utilisateur demande uniquement un résumé des côtés positifs de *Rudy Giuliani*, et par conséquent exclure du résumé toute phrase véhiculant une opinion négative. Les requêtes adoptent des structures assez régulières et comportent souvent des motifs communs. Apprendre à

8. Information atomique : information ne pouvant être découpée en plusieurs informations élémentaires.

les catégoriser peut donc se faire avec un nombre limité d'exemples et les campagnes d'évaluation précédentes fournissent un bon échantillon d'apprentissage. Comme les requêtes contiennent presque toujours un élément explicite indiquant la polarité, le classifieur commet très peu d'erreurs lors de l'analyse.

La classification des phrases issues des billets de blogs constitue un tout autre défi. En effet, la variabilité de la longueur des phrases, de leur structure et du vocabulaire utilisé constitue autant d'obstacles à la classification. De plus, de nombreuses phrases n'ont pas de polarité car elles ne mentionnent pas explicitement une opinion, ou n'utilisent pas de vocabulaire spécifique. Pour cette raison, nous avons adopté une stratégie plus simple mais potentiellement moins précise, qui consiste à classifier non les phrases seules, mais les billets de blogs. Ainsi, chaque phrase recevra la polarité du billet dont elle est issue. Nous avons adopté cette approche suite à l'étude minutieuse d'un ensemble de données typiques avant l'évaluation officielle. Nous avons observé les trois éléments suivants :

1) les billets de blogs sont très peu souvent nuancés. Ils véhiculent généralement une opinion quasi uniformément positive ou négative, ce qui nous a poussés à adopter cette stratégie de répercussion de la tendance générale d'un billet sur chaque phrase particulière ;

2) certaines phrases sont pertinentes bien qu'elles n'expriment pas en elles-mêmes une opinion. Leur affecter un score sur la base d'un contexte plus large (le billet de blog en l'occurrence) permet de résoudre partiellement ce problème ;

3) enfin, certaines phrases exigent l'analyse d'un contexte plus large pour en identifier la polarité.

Les phrases suivantes illustrent le deuxième point : « *At Carmax, the price is the price and when you want a car you go get one* » ou « *As Mayor of New York, Giuliani cut a plethora of taxes* ». Ces phrases expriment des faits plutôt que des opinions mais les billets contiennent par ailleurs des mots-clés indiquant que l'opinion exprimée est positive. On voit ici l'intérêt de disposer d'un système analysant finement le contenu des blogs, indépendamment du module d'analyse d'opinion.

Prenons par ailleurs la phrase : « *He only became popular because the excellent writers on that still-successful show played to the sole strengths Clooney has as an actor : his good looks and charm.* » On voit que cette phrase contient de nombreux éléments de nature positive (notamment *popular, excellent, good look, charm*). Cependant, l'analyse d'un contexte plus large peut corriger une analyse qui serait fausse si elle était limitée à cette seule phrase. Il n'est qu'à voir le titre du billet : « *George Clooney Is An Idiot* ». On est bien évidemment ici dans un cas extrême et tous les billets ne sont pas aussi caricaturaux, mais le genre même du blog pousse à avoir des avis extrêmement tranchés. L'analyse du contenu global des billets est donc beaucoup plus fiable que l'analyse de segments plus courts comme des phrases[9].

9. Notons néanmoins que des outils comme OpinionFinder (Riloff et Wiebe, 2003), permettent de distinguer les phrases objectives des phrases subjectives avec un certain succès. Les corpus

En général, les classifieurs pour l'analyse de sentiment sont assez fortement dépendants du domaine sur lequel ils ont été développés, en particulier ceux utilisant des techniques d'apprentissage supervisé comme le nôtre. Ceci est un problème quand on traite de données variées, portant sur plusieurs domaines comme les corpus de blogs. Nous avons tenté de pallier ce défaut en construisant un corpus d'entraînement hybride à partir de critiques de films[10] et de billets de blogs annotés par leur auteur selon l'humeur (Généreux et Evans, 2006). Les textes trop courts ont été écartés ; les billets de blogs annotés et les critiques de films restants ont permis de construire un corpus d'entraînement de 827 textes répartis comme suit :

– 206 textes « négatifs » obtenus à partir de critiques négatives[11] ou de billets de blogs annotés avec des humeurs négatives (triste, anxieux, etc.) ;

– 298 textes « positifs » obtenus à partir de critiques positives ou de billets de blogs annotés avec des humeurs positives (heureux, amoureux, etc.) ;

– 323 textes « neutres » obtenus à partir de critiques ou de billets de blogs ne faisant pas partie des deux autres catégories.

Ce corpus semble bien adapté dans la mesure où il mêle sentiments et opinions personnelles. Ces 827 critiques et billets de blogs ont donc été utilisés pour créer un classifieur SVM de type C-SVC (*regularized support vector classification*) avec un noyau RBF (*radial basis function*) à l'aide du logiciel LIBSVM[12]. Critiques et billets ont tous été étiquetés morphosyntaxiquement avec TreeTagger.

Pour éviter de prendre en compte les constructions négatives, nous avons omis tout trait apparaissant dans le voisinage de la négation (*not*). Il est en effet difficile de calculer la portée exacte de *not*, qui n'a d'ailleurs pas toujours une valeur négative en contexte. Ainsi, (Jia *et al.*, 2009) relèvent des exemples comme *not only*, *not just*, *not to mention*, qui ne doivent pas être analysés comme étant négatifs : la combinaison d'une particule négative et d'un terme polaire candidat dans un voisinage restreint peut donc être sans effet, ou avoir un effet positif ou négatif. Des approches complexes ont été proposées pour essayer de déterminer la portée de la particule négative et inverser la polarité d'un terme candidat se situant à l'intérieur de cette portée mais la précision de ces approches est restée limitée[13]. Pour éviter ces problèmes délicats, nous avons préférer ne pas tenir compte des termes polaires dans le voisinage (3 mots) de la particule de négation. Cette plage est assez grande pour capter, en moyenne, les effets de la particule sans diminuer de façon significative le nombre de traits disponibles pour

composés de données annotées au niveau de la phrase (Pang et Lee, 2005) permettent aussi d'élaborer des classifieurs s'attachant à ce niveau, si cela est nécessaire. Notre approche s'est révélée bien adaptée au contexte des blogs, mais elle ne conviendrait probablement pas pour d'autres cas nécessitant une analyse plus fine.

10. `http://www.cs.cornell.edu/People/pabo/movie-review-data/`

11. Il est possible de repérer la polarité de la critique en fonction du nombre d'étoiles attribuées au film.

12. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

13. « *The precision involving the negative word "not" is very low* » (Jia *et al.*, 2009).

le calcul général de la polarité. Le fait de travailler au niveau du document diminue de toute manière l'impact de cette simplification.

Afin de maximiser les performances du classifieur, trois paramètres sont particulièrement importants : les types de traits susceptibles de faire ressortir la polarité d'un texte, la valeur associée à chaque trait et enfin l'approche permettant de sélectionner les traits les plus discriminants.

1) Nous avons étudié trois groupes de traits : *i*) les catégories grammaticales (adjectifs, noms, verbes et adverbes), *ii*) les facettes linguistiques fonctionnelles (les facettes sont des regroupements fonctionnels de traits pertinents pour identifier des genres textuels, sur la base de l'étude de Généreux et Santini (2007)) et *iii*) les groupes de termes à connotation émotive, sur la base de WordNet-Affect (Valitutti, 2004) et du *Big-Six* (Ekman et Friesen, 1971).

2) L'éventail de valeurs que l'on peut associer à chaque trait est évidemment primordial. La méthode la plus simple est dite binaire, dans la mesure où seule la présence ou l'absence du trait est prise en compte. Une autre méthode simple consiste à utiliser la fréquence : le nombre d'apparitions du trait dans le document est directement pris en compte ou normalisé en fonction de la longueur de document.

3) Nous avons examiné trois approches pour la sélection de traits : *i*) selon le nombre de documents dans lesquels le trait apparaît (*document frequency*), *ii*) selon le gain d'information (*information gain*) ou *iii*) selon la statistique du χ^2.

Ces différentes possibilités ont été examinées dans diverses études antérieures (Généreux et Santini, 2007 ; Généreux *et al.*, 2008). Il en ressort que l'approche la plus efficace consiste à s'en tenir à des traits linguistiques simples (adjectifs, adverbes) avec des valeurs binaires, et sélectionnés sur la base du gain d'information. Nous avons donc paramétré notre classifieur sur cette base, en extrayant tous les adjectifs et adverbes, en indiquant si oui ou non le trait était présent dans un texte donné et en se limitant aux traits ayant une valeur de gain d'information résolument positive (pour un total de 150 traits). Parmi les adjectifs les plus discriminants repérés par la méthode, on trouve *new, great, recent, amazing, bouncy, previous, beautiful* mais aussi *bad, worst, viral*, ainsi que d'autres éléments dont le pouvoir discriminant n'est pas évident hors contexte, comme *online, previous* ou *commercial*, etc.

4.4. *Post-traitements*

4.4.1. *Gestion de la longueur du résumé*

La tâche « résumé d'opinion » autorise pour chaque résumé un nombre maximal de 7 000 caractères multiplié par le nombre de questions du sujet. Contrairement à la majorité des systèmes de résumé qui sélectionnent les phrases de manière itérative jusqu'à avoir atteint la taille maximale (Boudin et Torres-Moreno, 2007), nous ne pouvons relancer notre algorithme à plusieurs reprises dans la mesure où les classes risquent d'être largement différentes d'une itération à l'autre. Nous avons donc choisi

une stratégie plus simple, fondée sur la longueur moyenne des phrases. Le système calcule la longueur moyenne des phrases des blogs à résumer, puis extrait un nombre de phrases égal à :

nombre maximal de caractères / *nombre moyen de caractères par phrase*

Ce calcul permet de déterminer empiriquement le nombre de classes pour l'algorithme k-means.

4.4.2. *Ordonnancement des phrases*

Seules les phrases ayant une polarité compatible avec la requête sont retenues. Nous avons ensuite privilégié une stratégie d'ordonnancement par rapport au contenu informationnel de la requête : les phrases qui ont le plus d'éléments communs avec la requête sont placées en premier, celles qui en ont le moins sont placées à la fin.

5. Évaluation : participation à la campagne « résumé d'opinion » de TAC 2008

Nous présentons dans cette section les résultats obtenus par le système CBSEAS-Opinion lors de la campagne d'évaluation *Text Analysis Conference* (TAC) 2008.

5.1. *Présentation de la campagne « résumé d'opinion » de TAC 2008*

La tâche « résumé d'opinion » de TAC 2008 est inspirée d'un scénario applicatif réaliste, dans lequel un utilisateur souhaite connaître les opinions exprimées à propos d'une entité clairement identifiée (une personne, un produit, une organisation, un fait de société...). Même si les dépêches de presse peuvent parfois contenir ce type d'informations, celles-ci restent des sources secondaires formulées par des journalistes. Des médias comme les blogs permettent de disposer d'opinions non déformées. La tâche de TAC 2008 sur le résumé d'opinion (*Opinion Summarization Task*) visait l'analyse de telles données.

Dans ce scénario, l'utilisateur cherche des réponses à plusieurs questions spécifiques sur une entité donnée ou sur ses caractéristiques. Nous appellerons un jeu de questions à propos d'une même entité un « sujet ». L'utilisateur souhaite comme résultat final un résumé des réponses trouvées dans les blogs, où les informations redondantes ne sont pas répétées, mais éliminées. Ce résumé doit être bien organisé et sa lecture aisée. Le nombre de caractères en sortie ne doit pas dépasser 7 000 fois le nombre de questions. Par exemple, pour un sujet qui comporte trois questions, le résumé peut comprendre jusqu'à 21 000 caractères[14].

14. Cette limite en nombre de caractères peut paraître importante (surtout en comparaison du nombre limite de mots (100) pour la tâche « résumé et mise à jour » lors de la même campagne

Comme nous l'avons vu dans l'introduction, l'objectif de cette tâche est de fournir un résumé sur un thème (*le sujet*) en fournissant des réponses à des questions « orientées » vers une polarité d'opinion (positive ou négative). Par exemple, un jeu de questions issu de TAC 2008 était :

– sujet : *Rudy Giuliani presidential chances*

1) *What did American voters admire about Rudy Giuliani ? (+)*

2) *What qualities did not endear Rudy Giuliani ? (–)*

Seule une petite fraction des questions issues de la tâche « résumé d'opinion » de TAC 2008 n'était pas orientée vers une opinion positive ou négative, ou difficilement identifiable comme telle, comme ces deux questions en pourquoi (*why*) :

– sujet : *Criminalizing flag burning*

1) *Why do supporters want to make flag burning a crime ?*

2) *Why do opposers not want to make flag burning a crime ?*

La tâche « résumé d'opinion » consiste à coupler une analyse d'opinion avec un système de recherche d'information (les sujets et les questions ci-dessus proviennent d'une tâche de QR aussi organisée par NIST ; rappelons qu'il était éventuellement possible d'utiliser les résultats d'un système de QR mais que nous avons préféré participer à la sous-tâche reposant uniquement sur les résultats d'un système de recherche d'information, ceci rendant à notre avis l'application plus facilement portable et évaluable, dans la mesure où l'on peut alors plus facilement repérer la source des erreurs observées). Lors de TAC 2008, les sujets étaient au nombre de 25 et étaient constitués d'une à trois questions. Les systèmes devaient, pour chaque sujet, produire en sortie un texte qui résumait les réponses à toutes les questions de ce sujet. Les documents associés à chaque sujet sont censés être pertinents et sont fournis par NIST.

5.2. *Évaluation*

5.2.1. *Protocole*

Évaluer des résumés d'opinion est une tâche plus compliquée que d'évaluer des résumés « standard ». En effet, le juge, qu'il soit humain ou artificiel, doit noter non seulement la pertinence d'un résumé vis-à-vis d'une requête, mais également la polarité, qui doit être conforme à ce qu'a demandé l'utilisateur. Si l'évaluation automatique de la pertinence d'un résumé commence à être maîtrisée et fournit des résultats de plus en plus corrélés aux résultats manuels, repérer automatiquement si les opinions véhiculées par un résumé sont conformes n'est pas réaliste à l'heure actuelle.

TAC 2008) et en partie subjective ; NIST n'a pas donné d'explication pour avoir fixé une limite aussi haute.

NIST n'a donc pas pu automatiser complètement l'évaluation des résumés d'opinion lors de TAC 2008 : les évaluations étaient soit purement manuelles, guidées par une grille d'évaluation, soit fondées sur la méthode Pyramide. Cette méthode (Nenkova *et al.*, 2007) fonctionne de la manière suivante. À partir d'un ensemble de résumés de référence, une liste de SCU (*Single Content Units* ou unités de contenu) est établie. Ces SCU sont classés et pondérés selon leur fréquence d'apparition dans les résumés de référence (plus une unité est fréquente, plus son poids est important). La liste des SCU prend alors la forme d'une pyramide (une structure hiérarchisée) reflétant l'importance des SCU. L'étape suivante consiste à repérer dans les résumés à évaluer les SCU présents dans la pyramide. Un score est alors attribué aux résumés, correspondant à la somme des poids des SCU qu'il contient, divisée par la somme des poids des SCU des résumés de référence. Ce type d'évaluation, peu automatisable, a le défaut d'être extrêmement coûteux en temps d'expertise. De plus, apprendre à construire une pyramide est difficile, et la construction d'une pyramide de SCU est également coûteuse en temps. Cependant, cette méthode d'évaluation semble être la plus corrélée aux résultats d'évaluation subjective (sans protocole précis de notation) sur les résultats de la tâche de résumé des campagnes TAC (Dang et Owczarzak, 2008 ; Nenkova *et al.*, 2007)

Les résumés générés devaient répondre à toutes les questions d'un sujet donné. Par conséquent, les pyramides (constituées des éléments de réponses pertinents) ont été établies en tenant compte de la totalité des questions de chaque sujet. Cependant, pour chaque phrase d'un résumé, le lecteur doit être capable de déterminer, soit d'après la phrase elle-même soit d'après son contexte, à quelle question elle répond. Les éléments du résumé n'ont donc été associés à un élément de réponse (SCU) de la pyramide seulement si l'évaluateur a pu déterminer à quelle(s) question(s) ces éléments répondaient.

L'instanciation des SCU pour chaque pyramide s'est faite de la manière suivante : leur poids a été calculé à partir des jugements de dix évaluateurs, qui les ont notés en tant qu'« essentiels » (*vital*) ou « acceptables » (*okay*) vis-à-vis de la question posée. Le poids d'un SCU a alors été calculé en fonction du nombre d'évaluateurs qui l'ont marqué comme « essentiel » normalisé par le nombre de jugements « essentiels » reçus par le meilleur SCU (le plus essentiel) du sujet à résumer.

Une autre évaluation[15] correspondant à une grille de lecture et à une notation complètement subjective a été réalisée. Les évaluateurs ont noté de 1 à 10 les résumés selon les cinq critères suivants :

– grammaticalité : les résumés ne doivent pas comporter de phrases agrammaticales (fragments de phrases, composants manquants) qui rendent le texte difficile à lire ;

– non-redondance : les résumés de doivent pas inclure de répétitions inutiles. Cela inclut les phrases entières qui sont répétées aussi bien que des faits relatés plusieurs

15. En ce qui concerne l'évaluation des résumés, soulignons la contribution de Goulet (2007) qui va au-delà de la couverture des n-grammes et propose une terminologie adaptée au français.

fois ;

– structure et cohérence : les résumés doivent être bien structurés et organisés. Un résumé ne doit pas être seulement un amoncellement d'informations, mais doit être construit afin de constituer un corps d'informations cohérent. Si une personne ou une entité est mentionnée, la relation au reste du résumé doit être claire. Il doit être aisé d'identifier à qui ou à quoi les pronoms et les autres éléments non autoréférentiels font référence ;

– score total de lisibilité : le résumé est-il lisible ? Compréhensible ?

– *overall responsiveness* : les évaluateurs ont attribué un score global à chaque résumé, en fonction de la qualité de la réponse apportée au besoin informationnel exprimé par les questions. De façon concrète, les évaluateurs devaient se demander : « Si je me posais ces questions à propos de ce sujet, combien serais-je prêt à payer ? » (Dang et Owczarzak, 2008).

5.2.2. *Résultats*

Dans cette section, nous présentons nos résultats pour la tâche « *opinion Summarization* » de TAC 2008. Notre système n'utilisant pas les *snippets* (fragments de réponses apportés par les systèmes de QR), nous nous concentrons ici sur les systèmes qui n'ont pas utilisé cette ressource (nous situerons tout de même nos résultats par rapport à l'ensemble des systèmes, *cf.* tableau 2).

RunID	**Pyramide**	**Gram.**	**Non red.**	**Struct.**	**Lisib.**	***Over. Resp.***
Classement	5/19	3/19	3/19	2/19	2/19	10/19
CBSEAS	0,169	5,955	6,636	3,500	4,455	2,636
Moins bon score	0,101	3,545	4,364	2,045	2,636	1,682
Meilleur score	0,251	7,545	7,909	3,591	5,318	3,909

Tableau 1. *Résultats de TAC 2008* Opinion Summarization *: classement du système CBSEAS (systèmes sans* snippets*)*

RunID	**Pyramide**	**Gram.**	**Non red.**	**Struct.**	**Lisib.**	***Over. Resp.***
Classement	15/34	5/34	6/34	2/34	4/34	22/34
CBSEAS	0,169	5,955	6,636	3,500	4,455	2,636
Moins bon score	0,101	3,545	4,364	2,045	2,636	1,682
Meilleur score	0,489	7,545	8,045	3,591	5,318	5,773

Tableau 2. *Résultats de TAC 2008* Opinion Summarization *: classement du système CBSEAS (tous systèmes confondus)*

Le tableau 4 donne des exemples des résumés que notre système produit, en montrant les deux meilleurs résumés ainsi qu'un des moins bons résultats d'après l'évaluation Pyramide. Les résultats présentés dans le tableau 1 montrent le bon comportement de notre système de résumé vis-à-vis de cette tâche de résumé d'opinion. Le système se classe en effet cinquième sur dix-neuf systèmes d'après l'évaluation Pyramide, et entre la deuxième et la troisième place sur les évaluations de lisibilité. Il est à noter

RunID	Pyramide	Gram.	Non red.	Struct.	Lisib.	*Over. Resp.*
Moy. sans *snippets*	0,151	5,140	5,880	2,680	3,430	2,610
CBSEAS	0,169	5,955	6,636	3,500	4,455	2,636
Moy. avec *snippets*	0,312	5,450	5,910	2,870	3,880	4,040

Tableau 3. *Résultats de TAC 2008* Opinion Summarization *: comparaison des systèmes avec et sans* snippets

le comportement compétitif du système CBSEAS en ce qui concerne la structuration du résumé : le système est classé deuxième, à seulement 2,5 % du meilleur système. Cela montre l'efficacité de la méthode d'ordonnancement couplée au regroupement des phrases suivant leur polarité d'opinion. De plus, le système reste très bien classé face aux systèmes qui utilisent les *snippets* (*cf.* tableau 2), malgré l'avantage certain que cela leur a octroyé (*cf.* tableau 3).

5.3. *Discussion*

Comme on l'a vu dans la section précédente, l'évaluation effectuée par le NIST concerne le résultat final du système, à savoir les résumés produits. Le NIST n'offre pas d'évaluation spécifique de l'analyse d'opinion. On notera toutefois les scores moyens obtenus pour ce qui est de la satisfaction globale (*overall responsiveness*) : peut-être faut-il y lire un défaut de ce point de vue de notre analyse d'opinion ? Les bons scores obtenus en lisibilité et en grammaticalité montrent que les textes sont lisibles, mais manifestement ils ne répondent pas pleinement aux besoins des utilisateurs.

La stratégie que nous avons adoptée, à savoir reporter sur chaque phrase la polarité obtenue à partir d'un calcul portant sur la totalité du billet de blog, doit être discutée. Afin d'évaluer cette approche, une expérience complémentaire a été effectuée en utilisant OpinionFinder (Riloff et Wiebe, 2003) qui permet une analyse au niveau de la phrase. Ce système fonctionne en projetant sur les séquences à analyser des ressources prédéfinies. Quand on applique OpinionFinder à notre corpus, on voit qu'un nombre extrêmement important de phrases ne véhiculent pas d'opinion en tant que telle (ou, en tout cas, OpinionFinder est incapable de détecter une opinion et classe ces phrases comme neutres). Ce résultat rejoint l'observation qui avait été faite suite à une étude minutieuse du corpus : de nombreuses phrases ne véhiculent pas d'opinion, mais ces phrases peuvent malgré tout parfois être pertinentes. À côté de l'expression d'opinions au sens classique du terme, de nombreuses phrases sont en fait des justifications et des argumentations autour des jugements exprimés. Il est alors important d'analyser la polarité à un niveau plus global afin de pouvoir retenir dans le résumé produit ces phrases neutres mais importantes du point de vue de l'argumentation.

Il n'empêche qu'une analyse plus fine de l'opinion pourrait peut-être améliorer nos résultats. Ainsi, les deux meilleurs systèmes de la campagne d'évaluation semblent avoir apporté un soin particulier à la gestion des opinions. Murray *et al.* (2008) uti-

1) What motivated negative opinions regarding purchasing a car from CARMAX ? 2) What motivated positive opinions of CARMAX from car buyers ?
With Carmax you will generally always pay more than from going to a good used car dealer. Carmax did split the bill which made me happy. We bought it at Carmax, and I continue to have nothing bad to say about that company. Not sure if you have a Carmax near you, but I ve had 2 good buying experiences from them. At Carmax, the price is the price and when you want a car you go get one. Have to say that carmax rocks.

1) Why do people like Mythbusters ?
I watched a couple of episodes of MythBusters which was really entertaining. Mom spent a lot of money on getting the first season of the mythbusters. So the mythbusters did this with the dummy and, man it was so funny. My wife and I are huge Mythbusters fans. amhorach : I love Mythbusters. Mythbusters takes urban myths and proves or disproves them by recreating them in painstaking detail, like Alton, and like him, they use lots of zippy diagrams and helpful interviews with anthropologists and whenever an anthropologist gets a working gig, a tribeman gets a loincloth.

1) What reasons were given for liking Amita ? 2) Why did people like Megan ?
very solid story that felt real and i say that because i know i 've seen stuff like this on shows like primetime and 20/ 20, and it finally gave navi rawat she plays amita a reason to be in the credited cast. don and david stumbled on to this scene in the basement of an la hotel. at the fbi headquarters, amita is the only person that the victim will speak to. turns out the language the girl was speaking in was a dialect common to the area where amita's parents were from. i really liked how this episode forced amita to look at who she was and where she came from. originally, i think the supposition on the battery was as a method of torture. come on, that's nuts. does anyone know when/ if numb3rs is coming out on dvd. how many shows can drag on and on. same scenario ... case opens, call in charlie, larry sneaks up on charlie either in the garage or at school, small mention/ scene with how many shows can drag on and on. she had been gone on a trip around several asian countries for the latter part of october, so we have n't really been e-mailing or text messaging. so she updated me on a lot of shows and we talked about our recent trips. even if i have to sail on a boat to get there. hopefully did n't do too bad on any of them. don comforting terry ... charlie and amita, passing pens michelle and i were probably the only ones who noticed that, as it was n't the focus of the scene and they were just off on the side ... according to my sister, the junior reshuffle is n't so bad. i know the education system there is alot tougher it would seem then ours from example but really it will be to your benefit later on.

Tableau 4. *Exemples de résumés générés par CBSEAS pour la tâche « Opinion Summarization » de TAC 2008 (*Mythbusters *et* Numb3rs *sont deux séries télévisées ;* Amita *et* Megan *sont deux personnages de la série Numb3rs)*

lisent une combinaison de deux scores de polarité, l'un issu d'une analyse à base de règles apprises automatiquement sur corpus (selon une stratégie identique à celle décrite dans (Riloff et Wiebe, 2003)), l'autre fondé sur des informations dictionnairiques (Taboada *et al.*, 2008). Li *et al.* (2008) ont quant à eux utilisé une approche spécifique aux blogs : la polarité d'une phrase est donnée par la somme de sa polarité et de celle de son paragraphe ; les phrases issues des commentaires (commentaires des lecteurs attachés au billet de blog) sont favorisées par rapport à celles issues des billets initiaux, car elles paraissent plus diversifiées. Ces deux systèmes employant par ailleurs des techniques de résumé traditionnelles nous permettent de remarquer qu'une bonne gestion de l'analyse d'opinion peut jouer un rôle prépondérant.

On voit enfin, à la lecture des résumés produits (*cf.* tableau 4), que les phrases sélectionnées dans les blogs ne forment pas un tout aussi cohérent que le résultat d'un résumé de dépêche (même si les opinions sont regroupées suivant leur polarité). Le style des blogs est personnel, divers, avec des niveaux de langue très fortement variables. L'emploi de la première personne rend le résultat plus douteux que pour un résumé de dépêche : il faut bien reconnaître que les résumés donnent l'impression d'un amalgame d'opinions et de réflexions variées sans grande unité, malgré les bons scores obtenus lors de la campagne TAC 2008. Il serait sans doute préférable, dans ce cas, de présenter explicitement les différentes opinions en fonction des différents blogs (c'est-à-dire présenter des extraits plus que des résumés). TAC 2008 ne poussait pas dans cette voie, mais ce type de sorties serait facilement envisageable à partir des analyses mises en place[16].

6. Conclusion

Cet article a présenté la mise au point d'un système de résumé multidocument orienté vers l'analyse d'opinion. Notre système a obtenu des résultats compétitifs à TAC 2008 en se classant dans les cinq premiers de sa catégorie dans l'évaluation manuelle de type Pyramide et dans les évaluations de lisibilité. Les résultats de TAC 2008 mettent aussi en avant la qualité de notre approche d'ordonnancement des phrases, qui offre à l'utilisateur un texte plus organisé et une lecture plus fluide que ce que propose la grande majorité des systèmes concurrents.

La stratégie de report de la polarité d'opinion calculée au niveau du billet de blog sur chaque phrase de celui-ci est simple mais ne semble pas déraisonnable quand on prend en compte la dimension textuelle des billets de blogs. Nous avons montré que ceux-ci comportent de nombreuses phrases ne véhiculant pas directement d'opinion mais des informations ou des arguments intéressants malgré tout pour le type de résumé visé. L'évaluation faite par le NIST ne permet pas de mesurer finement l'apport

16. Ceci serait aussi plus satisfaisant sur le plan « éthique », dans la mesure où il semble important de garder un lien vers les blogs où les opinions ont été exprimées et, surtout, dans la mesure où il faut être extrêmement prudent et éviter au maximum de suggérer une interprétation à partir de séquences de textes hors contexte.

des différents modules du système de résumé ; cette évaluation est du coup difficile à faire *a posteriori*. Il est donc difficile de tirer des conclusions solides quant à l'adéquation de l'approche de l'analyse de l'opinion par rapport à la tâche.

Les perspectives concernent d'une part l'amélioration de la présentation du texte, afin d'éviter d'agréger trop sommairement des opinions émanant de différentes sources. Une autre piste vise à simplifier les phrases afin d'obtenir des résumés encore plus condensés en s'appuyant sur une analyse linguistique plus poussée que celle que nous avons mise en œuvre ici. Enfin, pour ce qui est des résumés orientés par une requête comme dans le cas de TAC 2008, un traitement plus fin des questions en entrée pourrait s'avérer utile (analyse syntaxique de la requête, extension au moyen d'un réseau sémantique, etc.).

Remerciements

Nous tenons à remercier les relecteurs anonymes de la revue TAL pour leurs remarques pertinentes qui nous ont permis de grandement améliorer la qualité de l'article. Nous remercions également Béatrice Pelletier pour sa relecture attentive.

Ce travail a été effectué alors que les auteurs travaillaient au Laboratoire d'Informatique de Paris-Nord. Il a été en partie financé par le projet *INFOM@GIC*, dans le cadre du pôle de compétitivité *CAP DIGITAL*.

7. Bibliographie

Asher N., Benamara F., Mathieu Y., « Distilling opinion in discourse : A preliminary study », *Proceedings of COLING, companion volume*, Manchester, 2008.

Baccianella S., Esuli A., Sebastiani F., « SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining », *Proceedings of the conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

Bethard S., Yu H., Thornton A., Hatzivassiloglou V., Jurafsky D., « Automatic extraction of opinion propositions and their holders. », *Working Notes of the AAAI Spring Symposium on Exploring Attitude and Aect in Text : Theories and Applications*, 2004.

Bossard A., Généreux M., Poibeau T., « Description of the LIPN Systems at TAC2008 : Summarizing Information and Opinions », *Proceedings of the Text Analysis Conference*, NIST, Gaithersburg, 2008.

Boudin F., Torres-Moreno J. M., « NEO-CORTEX : A Performant User-Oriented Multi-Document Summarization System », *in* A. F. Gelbukh (ed.), *CICLing*, vol. 4394 of *Lecture Notes in Computer Science*, Springer, p. 551-562, 2007.

Boudin F., Torres-Moreno J.-M., El-Bèze M., « A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization », *COLING Conference*, Manchester,UK, p. 21-24, August, 2008.

Carbonell J., Goldstein J., « The use of MMR, diversity-based reranking for reordering documents and producing summaries », *SIGIR 1998 : Proceedings of the 21st Annual Internatio-*

nal ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Melbourne, Australie, August, 1998.

Cardie C., Wiebe J., Wilson T., Litman D., « Combining low-level and summary representations of opinions for multi-perspective question answering », *In Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, p. 20-27, 2003.

Dang H. T., Owczarzak K., « Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks », *Proceedings of the TAC 2008 Workshop*, vol. Notebook Papers and Results, 2008.

Dey L., Haque K., « Opinion mining from noisy text data », *AND 08 : Proceedings of the second workshop on Analytics for noisy unstructured text data*, NewYork, 2008.

Edmundson H. P., « New Methods in Automatic Extracting », *Journal of the Association for Computing Machinery*, 1969.

Ekman P., Friesen W., « Constants across cultures in the face and emotion », *Journal of Personality and Social Psychology*, vol. 17, p. 124-129, 1971.

Erkan G., Radev D. R., « LexRank : Graph-based Centrality as Salience in Text Summarization », *Journal of Artificial Intelligence Research (JAIR)*, 2004.

Esuli A., Sebastiani F., « SentiWordNet : A Publicly Available Lexical Resource for Opinion Mining », *Proceedings of Language Resources and Evaluation*, 2006.

Forgy E., « Cluster analysis of multivariate data : Efficiency vs. interpretability of classifications », *Biometrics*, vol. 21, p. 768-769, 1965.

Fry E. B., Kress J. E., Fountoukidis D. L., *The Reading Teachers Book of Lists*, 4th edn, Jossey-Bass, Hoboken, 2000.

Gale W., Church K. W., Yarowsky D., « One Sense Per Discourse », *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, New York, p. 233-237, 1992.

Généreux M., « Summarizing a Blog Search Engine Hits », *Proceedings of the WWW 2009 Conference, Workshop on Web Search Result Summarization and Presentation*, 2009.

Généreux M., Evans R., « Towards a validated model for affective classification of texts », *Proceedings of the Workshop on Sentiment and Subjectivity in Text (SST'06)*, Association for Computational Linguistics, Morristown, NJ, USA, p. 55-62, 2006.

Généreux M., Poibeau T., Koppel M., « Sentiment analysis using automatically labelled financial news », *LREC 2008 Workshop on Sentiment Analysis : Emotion, Metaphor, Ontology and Terminology*, Marrakech, Morocco, May, 2008.

Généreux M., Santini M., « Exploring the Use of Linguistic Features in Sentiment Analysis », *Corpus Linguistics 2007*, Birmingham, UK, 2007.

Goulet M.-J., « Terminologie et paramètres expérimentaux pour l'évaluation des résumés automatiques », *Traitement Automatique des Langues*, 2007.

Hu M., Liu B., « Mining and summarizing customer reviews », *KDD'04 : Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, p. 168-177, 2004.

Hutchins W. J., Somers H. L., *An Introduction to Machine Translation*, Academic Press, London, 1992.

Jia L., Yu C., Meng W., « The effect of negation on sentiment analysis and retrieval effectiveness », *Proceeding of the 18th ACM conference on Information and knowledge management (CIKM)*, p. 1827-1830, 2009.

Joachims T., *Text Categorization With Support Vector Machines : Learning with many relevant features*, University Dortmund, 1997.

Kao H.-A., Chen H.-H., « Comment Extraction from Blog Posts and Its Applications to Opinion Mining », *Proceedings of the conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

Kim H. D., Zhai C. X., « Generating Comparative Summaries of Contradictory Opinions in Text », *Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM'09)*, 2009.

Li W., Ouyand Y., Hu Y., Wei F., « PolyU at TAC 2008 », *Proceedings of 2008 Text Analysis Conference*, 2008.

Likas A., Vlassis N., Likas A., Vlassis N., Verbeek J., « The Global K-Means Clustering Algorithm », *Pattern Recognition*, vol. 36, p. 451-461, 2001.

Luhn H. P., « The Automatic Creation of Literature Abstracts », *IBM Journal of Research Development*, vol. 2, n° 2, p. 159-165, 1958.

MacQueen J., « Some methods for classification and analysis of multivariate observations », *in* L. M. L. Cam, J. Neyman (eds), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Statistics, University of California Press, 1967.

Mani I., Maybury M. T., *Advances in Automatic Text Summarization*, MIT Press, 1999.

Marcu D., *The Theory and Practice of Discourse Parsing and Summarization*, The MIT Press, Cambridge, 2000.

Mathieu Y., « Annotation of Emotions and Feelings in Texts », *Conference on Affective Computing and intelligent Interaction*, p. 350-357, 2005.

Murray G., Joty S., Carenini G., Ng R., « The University of British Columbia at TAC 2008 », *Proceedings of the 2008 Text Analysis Conference*, 2008.

Nenkova A., Passonneau R., McKeown K., « The Pyramid Method : Incorporating Human Content Selection Variation in Summarization Evaluation », *ACM Transactions on Speech and Language Processing*, vol. 4, n° 2, p. 1-23, 2007.

Pak A., Paroubek P., « Twitter as a Corpus for Sentiment Analysis and Opinion Mining », *Proceedings of the conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

Pang B., Lee L., « Seeing stars : Exploiting class relationships for sentiment categorization with respect to rating scales », *Proceedings of the ACL*, 2005.

Radev D., Allison T., Blair-Goldensohn S., Blitzer J., Çelebi A., Dimitrov S., Drabek E., Hakim A., Lam W., Liu D., Otterbacher J., Qi H., Saggion H., Teufel S., Topper M., Winkel A., Zhu Z., « MEAD - a platform for multidocument multilingual text summarization », *Language Resource and Evaluation conference*, 2004.

Radev D. R., Blair-goldensohn S., Zhang Z., « Experiments in single and multidocument summarization using MEAD », *In Proceedings of the First Document Understanding Conference*, 2001.

Riloff E., Wiebe J., « Learning Extraction Patterns for Subjective Expressions », *Conference on Empirical Methods in Natural Language Processing (EMNLP-03). ACL SIGDAT.*, 2003.

Saggion H., « Topic-Based Summarization at DUC 2005 », *Proceedings of DUC 2005 Document Understanding Conference*, vol. Document Understanding Workshop, October, 2005.

Salton G., Buckley C., « Term-Weighting Approaches in Automatic Text Retrieval », *Information Processing and Management : an International Journal*, vol. 24, p. 513-523, 1988.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49, 1994.

Seki Y., « Summarization Focusing on Polarity or Opinion Fragments in Blogs », *Proceedings of the 2008 Text Analysis Conference*, 2008.

Spärck Jones K., Willett P., *Readings in Information Retrieval*, The Morgan Kaufmann Series in Multimedia Information and Systems, San Francisco, 1997.

Strapparava C., Valitutti A., « WordNet-Affect : an affective extension of WordNet », *Proceedings of Language Resources and Evaluation*, p. 1083-1086, 2004.

Taboada M., Voll K., Brooke J., « Extracting sentiment as a function of discourse structure and topicality », *Tech. report*, Simon Fraser University, 2008.

Turney P. D., « Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews », *40th Annual Meeting of the Association for Computational Linguistics, Philadelphia*, 2002.

Valitutti R., « WordNet-Affect : an Affective Extension of WordNet », *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, p. 1083-1086, 2004.

Vernier M., Monceaux L., « Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques », *Traitement Automatique des Langues*, 2007.

Wiebe J., Wilson T., Bell M., « Identifying Collocations for Recognizing Opinions », *ACL 01 Workshop on Collocation*, Toulouse, France, July, 2001.

Zhang W., Yu C., Meng W., « Opinion retrieval from blogs », *CIKM '07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, NewYork, 2007.