Résumé automatique guidé de textes : État de l'art et perspectives

Salima Lamsiyah[1] Saïd Ouatik El Alaoui[1] Bernard Espinasse[2]
(1) LIM, Faculté des Sciences Dhar El Mahraz, Université Sidi Mohamed Ben Abdellah, Fès, Maroc
(2) LSIS-UMR CNRS, Aix-Marseille Université, Marseille, France
Salima.lamsiyah@usmba.ac.ma, said.elalaouiouatik@usmba.ac.ma,
bernard.espinasse@lis-lab.fr

RÉSUMÉ

Les systèmes de résumé automatique de textes (SRAT) consistent à produire une représentation condensée et pertinente à partir d'un ou de plusieurs documents textuels. La majorité des SRAT sont basés sur des approches extractives. La tendance actuelle consiste à s'orienter vers les approches abstractives. Dans ce contexte, le résumé guidé défini par la campagne d'évaluation internationale TAC (Text Analysis Conference) en 2010, vise à encourager la recherche sur ce type d'approche, en se basant sur des techniques d'analyse en profondeur de textes. Dans ce papier, nous nous penchons sur le résumé automatique guidé de textes. Dans un premier temps, nous définissons les différentes caractéristiques et contraintes liées à cette tâche. Ensuite, nous dressons un état de l'art des principaux systèmes existants en mettant l'accent sur les travaux les plus récents, et en les classifiant selon les approches adoptées, les techniques utilisées, et leurs évaluations sur des corpus de références. Enfin, nous proposons les grandes étapes d'une méthode spécifique devant permettre le développement d'un nouveau type de systèmes de résumé guidé.

ABSTRACT

Guided Summarization : State-of-the-art and perspectives

Automatic text summarization (ATS) aims to produce from one or more texts, a summary that represents the most relevant information included in the original textual sources. Most existing ATS are mainly the extraction-based systems ; however, the trend today is to make a move toward abstraction-based systems. In 2010, the Text Analysis Conference (TAC) campaign defined the guided summarization task as a recent type of ATS, aims to encourage a deeper linguistic analysis of the source documents instead of relying only on classical extractive approaches. In this work, we provide an introduction to guided summarization task, by defining the different characteristics and constraints related to this task, and by reviewing the details of different guided summarization systems developed so far. We also classify these systems according to the adopted approaches, techniques used, and evaluations on reference corpus. Finally, we propose the main steps of a specific method that will allow the development of a new type of guided summary systems.

MOTS-CLÉS : Résumé automatique de textes, résumé guidé, approche extractive, approche abstractive, traitement automatique de la langue naturelle, extraction d'information.

KEYWORDS: Automatic text summarization, guided summarization, extraction-based approach, abstraction-based approach, natural language processing, information extraction.

1 Introduction

De nos jours, l'information textuelle en format numérique est abondante dans le web, elle représente une masse de 80% de l'information qui y circule. Dans la plupart des cas, cette immense quantité est non structurée : elle n'est pas sous forme de bases de données classiques, mais sous un format de texte libre nécessitant le besoin de concevoir et de développer des outils pertinents pour que l'utilisateur puisse accéder aux informations pertinentes.

Le résumé automatique de textes est un de ces outils, en condensant les textes de façon pertinente, le résumé automatique pourra être une solution efficace et éprouvée pour traiter cette masse grandissante d'informations.

Le résumé automatique de textes, apparu vers la fin des années 1950 (Luhn, 1958), a connu un fort renouveau ces dernières années. Produire automatiquement un résumé pertinent et de qualité nécessite de condenser le ou les documents originaux tout en minimisant la redondance, et en maximisant la cohérence et la cohésion. *La cohérence* est l'absence de contradictions et de la redondance dans l'enchainement des phrases d'un document. Pour toute partie d'un texte cohérent, il existe une fonction, une raison plausible à sa présence. *La cohésion* est un moyen de lier ensemble les différentes parties du texte, elle est assurée par l'utilisation de termes sémantiquement liés, co-références, ellipses et conjonctions. La cohésion se situe au niveau linguistique de la phrase, alors que la cohérence est située au niveau supérieur de la sémantique. De façon générale les systèmes de résumé automatique de textes (SRAT) sont confrontés à une difficulté majeure qui est l'absence d'un étalon-or unique (Gold standard) que les SART pourraient suivre, le résumé est guidé par une vague notion de l'importance des faits mentionnés dans les documents sources, ce qui est très subjectif et dépendant du contenu.

Une autre difficulté est liée à l'utilisation exclusive des approches extractives qui sont devenues le paradigme dominant du développement de SRAT. Bien que ces approches soient relativement simples, faciles à mettre en œuvre et efficaces en extraction des phrases pertinentes, elles sont loin de produire des résumés optimaux. En effet des problèmes de cohésion, la cohérence et de résolution d'anaphores empêchent les SRAT basés sur ce type d'approche de produire des bons résumés en termes du contenu et de la qualité linguistique. Les expériences HexTAC (Genest *et al.*, 2009) et d'autres études (Cremmins, 1993, 1996) ont montré que même le meilleur mécanisme contenu-sélection utilisé par les êtres humains est incapable de produire de bons résumés s'il se limite à assembler un ensemble de phrases prises hors de leurs contextes. En conséquence, la recherche en SRAT devrait s'orienter plus vers l'abstraction que vers l'extraction. L'une des thématiques récentes accompagnant cette tendance est le résumé guidé. Lancé dans la campagne internationale d'évaluation TAC'2010, le résumé guidé peut être considéré comme des résumés multi-documents dont les contenus sont déterminés par les besoins et les préférences des utilisateurs.

Le reste de cet article est organisé comme suit : la section 2 présente les différents types et approches de résumés automatiques de textes, la section 3 est dédiée au résumé guidé, sa définition, une architecture fonctionnelle générique, et les spécificités qui le distinguent des autres résumés. La section 4 présente brièvement plusieurs systèmes de résumé guidé et les compare selon différents critères. Dans la section 5 nous concluons et présentons différentes perspectives de recherche liées à l'amélioration des systèmes automatiques de résumé guidé.

2 Résumé automatique de textes

Bien que cet article s'intéresse au résumé guidé et à ses approches, nous présentons un aperçu des autres types de résumés, ainsi que les principales stratégies pour produire un résumé automatique.

2.1 Types de résumés automatiques de textes

Il existe plusieurs types de résumés de textes, du fait que l'on dispose de différents types et sources documentaires, et que le besoin en information diffère d'un utilisateur à un autre. Différentes taxonomies sont proposées pour les classifier (Sparck, 1998; Nenkova & McKeown, 2012). Nous présentons l'une des plus connues dans la littérature et qui classifie les résumés selon quatre critères (Lloret & Palomar, 2012) : l'entrée du SRAT, l'objectif, la sortie et la langue. En se basant sur le premier facteur, nous distinguons les SRAT *mono-documents* et les SRAT *multi-documents*. Les premiers produisent des résumés à partir d'un seul document alors que les deuxièmes génèrent des résumés pour un ensemble de documents et portant souvent sur une thématique bien précise.

Selon le critère objectif du SRAT, on distingue plusieurs types : le résumé indicatif, le résumé informatif, le résumé générique, le résumé orienté et le résumé de mise à jour. Le *résumé indicatif* a pour objectif d'aider le lecteur à agir sur sa décision à consulter ou pas un document, en lui indiquant les thématiques abordées et développées dans le document source, sans considérer les détails. Le *résumé informatif* a pour objectif principal de renseigner le lecteur sur les principales informations quantitatives et qualitatives, il est considéré comme une version abrégée, conservant l'organisation générale du document source. Le *résumé générique* résume le document sans prendre en compte les besoins en information des utilisateurs, par contre, le *résumé orienté* a pour objectif de ne résumer que les informations qui répondent à une requête de l'utilisateur. Le *résumé de mise à jour* se contente de fournir un résumé sous l'hypothèse que l'utilisateur a déjà des connaissances sur la thématique et qui n'a besoin que des nouveautés importantes, tout en évitant la redondance de l'information (Li *et al.*, 2009).

Selon le facteur de la langue, nous envisageons trois types des SRAT : monolingues, multilingues, et cross-lingues. Pour les *monolingues*, la source et le résumé sont écrits dans la même langue. Si le système peut traiter plusieurs langues, et produire des résumés dans la même langue que celle du document d'entrée, nous aurions un système *multilingue*. Si le résumé est en anglais, et que les documents originaux sont dans une autre langue, le SRAT est dit *cross-lingue*.

Finalement, en se basant sur la sortie du SRAT, nous distinguons deux grands types de résumés : le résumé extractif et le résumé abstractif. Le *résumé extractif* est généré par la sélection des phrases pertinentes et informatives telles qu'elles apparaissent dans les documents sources. Alors que le *résumé abstractif* (résumé par compréhension) se base sur des techniques qui utilisent une analyse en profondeur de textes pour produire de nouvelles phrases grammaticalement correctes, concises, cohérentes, devant donner un résultat proche d'un résumé humain.

La diversité des sources d'information contenues dans le web a poussé la communauté des chercheurs en résumé automatique de textes à ajouter un autre critère de classification des SRAT : le genre des documents sources. Selon ce facteur, il est également possible d'envisager d'autres types de SRAT à savoir : résumé d'articles de presse, résumé d'un domaine spécialisé (biomédical, droit, etc.), résumé des documents narratifs et de textes littéraires, résumé des pages web, résumé des conversations email, etc. Enfin, il est à noter que ces types de résumés ne sont pas indépendants les uns des autres. Un

résumé textuel qui est rattaché à un type de résumé particulier peut aussi être rattaché à un autre, dans la mesure où il répond à toutes les conditions assurant les fonctions de l'autre type.

2.2 Approches du résumé automatique

Pour générer un résumé automatique de textes, plusieurs méthodes et techniques sont proposées. Principalement, deux grands types d'approches s'opposent : l'approche par extraction et l'approche par abstraction. Plus récemment de nouvelles approches sont apparues, considérées comme semi-extractives mettant en œuvre des techniques de compression, de fusion et de division de phrases. Dans cette section, nous présentons brièvement ces approches.

2.2.1 Approche extractive

Les approches extractives cherchent à repérer et à extraire les segments textuels les plus pertinents pour constituer un résumé. Généralement, le processus de la génération d'un résumé par extraction comporte 4 étapes.

— *Prétraitement* (Analyse et représentation des documents) : les documents sources sont sous une forme non structurée ; cette étape permet de prétraiter ces documents pour les représenter de manière structurée. Le prétraitement implique généralement certaines techniques du TALN, notamment la segmentation de phrases, la tokenisation, la lemmatisation/stemming, la reconnaissance des entités nommées, la résolution de co-référence. Une fois que le prétraitement est terminé, une représentation des documents sources est requise, et elle consiste généralement en la représentation de chaque document par un vecteur, afin de le rendre exploitable par les algorithmes.

— *Pondération de phrases* : cette étape est cruciale pour un SRAT par extraction. En se basant sur la représentation déjà créée dans la première étape et sur des caractéristiques de phrases (Oliveira *et al.*, 2016a), cette phase consiste à assigner à chaque phrase un score indiquant sa pertinence. Plusieurs méthodes sont développées pour cette tâche. Nous citons les plus répandues dans la littérature : méthodes statistiques (Ko & Seo, 2008), méthodes basées sur les graphes (Mihalcea, 2004; Erkan & Radev, 2011; Baralis *et al.*, 2013), méthodes utilisant l'apprentissage automatique (Fattah, 2014; Yang *et al.*, 2014), méthodes utilisant les réseaux de neurones (Nallapati *et al.*, 2017), ainsi que d'autres récentes méthodes.

— *Sélection de phrases* : après avoir calculé les scores des phrases, nous sélectionnons celles ayant un score élevé pour générer un résumé. L'un des problèmes les plus importants de cette étape est d'éviter la redondance, et notamment pour les résumés multi-documents. Pour cela, plusieurs méthodes ont été introduites telles que MMR (Maximum Marginal Relevance) (Carbonell & Goldstein, 1998) et ILP (Integer Linear Programming) (Oliveira *et al.*, 2016b).

— *Génération du résumé* : généralement le système combine les phrases sélectionnées dans l'étape précédente telles qu'elles apparaissent pour générer un résumé.

2.2.2 Approche semi-extractive

La compression, la fusion et la division de phrases sont des axes de recherche relativement récents dans le résumé automatique de textes caractérisant les approches semi-extractives. Ces tâches de traitement des phrases permettent un certain nombre d'améliorations, notamment la réduction de la redondance, et la création de résumés plus proches des résumés abstractifs. *Les approches compressives* consistent à transformer une phrase pertinente en une phrase grammaticalement plus courte qui conserve l'information importante (Knight & Marcu, 2000; Zajic *et al.*, 2008; Torres-Moreno, 2014). *La fusion de phrases* consiste à générer une phrase simple, grammaticalement correcte à partir d'un ensemble de phrases connexes, et qui préserve les informations importantes de cet ensemble. Cette phrase n'est pas obligatoirement contenue dans cet ensemble (Tzouridis *et al.*, 2014; Torres-Moreno, 2014). *La division de phrases* est une nouvelle approche semi-extractive proposée par (Genest & Lapalme, 2011). Cette approche consiste tout d'abord à trouver des Information Items (InIts), qui sont définis comme étant les plus petits éléments d'information cohérents dans une phrase ou dans un texte. Puis, sélectionner ceux qui répondent au besoin d'information de l'utilisateur. Enfin, générer un résumé qui contient les InIts les plus pertinents.

2.2.3 Approche abstractive

Les méthodes de cette approche sont apparues vers la fin des années 1970. Elles s'inspirent principalement du domaine de l'intelligence artificielle et de la psychologie cognitive, et elles cherchent à produire des résumés avec une qualité linguistique comparable à celle des résumés produits par les êtres humains. Généralement, on distingue trois familles de méthodes pour l'approche abstractive (Andhale & Bewoor, 2016) : (i) des méthodes qui traduisent les informations importantes contenues dans les documents sources en des schémas cognitifs tels que les ontologies, les patrons, les graphes, (ii) les méthodes se basant sur la représentation sémantique des documents, et enfin (iii) les méthodes utilisant les réseaux de neurones dans le cadre de l'apprentissage profond (Deep Learning) (Nallapati *et al.*, 2016; Rush *et al.*, 2015).

2.3 Evaluation de systèmes de résumé automatique

L'évaluation de la qualité des résumés automatiques reste toujours une tâche extrêmement subjective et difficile, à laquelle la communauté scientifique a répondu avec des solutions partielles. Les méthodes d'évaluation, telles que la précision et le rappel, qui sont très utilisées dans les systèmes de recherche d'information, ne sont pas vraiment adaptées à cette tâche, vu que les entrées et les sorties des systèmes de résumé automatique sont des textes en langage naturel difficiles à comparer. Les méthodes d'évaluation peuvent être classées en deux types : d'une part les méthodes *intrinsèques* qui évaluent le résumé lui-même en fournissant des mesures automatiques ou semi-automatiques de l'informativité et d'autre part les méthodes *extrinsèques* qui mesurent la qualité du résumé à travers d'autres applications de la fouille de textes.

En ce qui concerne les méthodes *intrinsèques*, citons la méthode *ROUGE* (Lin, 2004) qui est la méthode la plus utilisée, elle est fondée sur la comparaison automatique de n-grammes entre un ou plusieurs résumés de référence et un résumé à évaluer. Il en existe plusieurs variantes, notamment ROUGE-n, ROUGE-SUn et ROUGE-L. La méthode *Pyramide* (Nenkova & Passonneau, 2004) est une autre méthode intrinsèque d'évaluation, mais semi-automatique, ayant pour objectif de surmonter

le problème de sémantique non abordé par ROUGE. Citons aussi la mesure *Responsiveness* qui évalue manuellement le résumé de point de vue du contenu et de la qualité linguistique. Pour les méthodes *extrinsèques*, mesurant la qualité du résumé à travers d'autres applications de la fouille de textes, telles que les systèmes de recherche d'information, la catégorisation de texte et les systèmes Questions-Réponses (Q/R), et qui sont spécifiques à la nature de ces applications.

Pour conclure sur les méthodes d'évaluation, notons que l'évaluation des résumés est une problématique à part entière, à laquelle la campagne TAC a proposé la tâche AESOP (Automatically Evaluating Summaries Of Peers) pour encourager le développement des méthodes automatiques d'évaluation des résumés.

3 Le résumé automatique guidé de textes

Cette section est consacrée à la description du résumé automatique guidé de textes, en se concentrant sur ses spécificités et les contraintes qu'il impose, ainsi que sur l'architecture fonctionnelle générique pouvant lui être associée.

3.1 Définition

La conférence Document Understanding Conference (DUC-2001/2007), et sa remplaçante Text Analysis Conference (TAC) organisées par le NIST (National Institute of Standards and Technology), ont présenté plusieurs types du résumé automatique de textes tels que le résumé orienté, le résumé multi-documents, le résumé de mise à jour. En 2010, la campagne internationale d'évaluation TAC a lancé une nouvelle tâche intéressante et qui représente un grand défi : le résumé automatique guidé de textes. Ce dernier propose un changement significatif vers des résultats orientés sémantiquement tout en favorisant le traitement profond du langage naturel, l'extraction d'information spécifique à un domaine, l'utilisation des ontologies, etc. Et cela dans l'optique d'encourager le passage vers la génération des résumés par abstraction.

La campagne TAC'2010/2011 définit le résumé guidé comme étant un résumé multi-documents de 100 mots obtenu à partir d'un ensemble de 10 articles de presse qui portent sur une thématique précise et appartenant à une catégorie préalablement définie. En l'occurrence, il s'agit d'articles relatifs à des attaques terroristes. Cinq *catégories* ont été sélectionnées et chaque catégorie comporte une liste spécifique *d'aspects* qui sont définis conformément aux thématiques des documents. Selon Jin *et al.* (2011), un aspect est défini comme un thème sémantique représentant un attribut important des entités trouvées dans la collection des documents. Les catégories et leurs aspects ont été développés à partir des modèles des résumés tirés des campagnes DUC/TAC déjà passées. Par exemple, les aspectes de la catégorie ATTACKS sont : WHAT (What Happened), WHEN (date, time, ...), WHERE (location), PERPETRATORS (individuals or groups responsible for the attack), WHY (reasons), WHO AFFECTED (casualities), DAMAGES (caused by the attack) and COUNTERMEASURES (rescues efforts, preventive effort, ...). Le résumé doit couvrir tous ces aspects, comme il peut également contenir d'autres informations pertinentes.

Un système de génération de résumé guidé nécessite également un composant du résumé de mise à jour destiné à produire un résumé sous l'hypothèse que l'utilisateur a déjà lu le résumé des 10 premiers articles, et il n'a besoin que des dernières nouvelles. Alors, le résumé guidé répond à deux

demandes émergentes de traitement de l'information : les exigences en termes d'aspects spécifiques et de temps.

3.2 Architecture fonctionnelle générique

Bien que les détails d'implémentation des systèmes automatiques de résumé guidé soient différents les uns des autres, cependant une architecture fonctionnelle générique illustrée à la figure 1 peut être définie pour les systèmes.

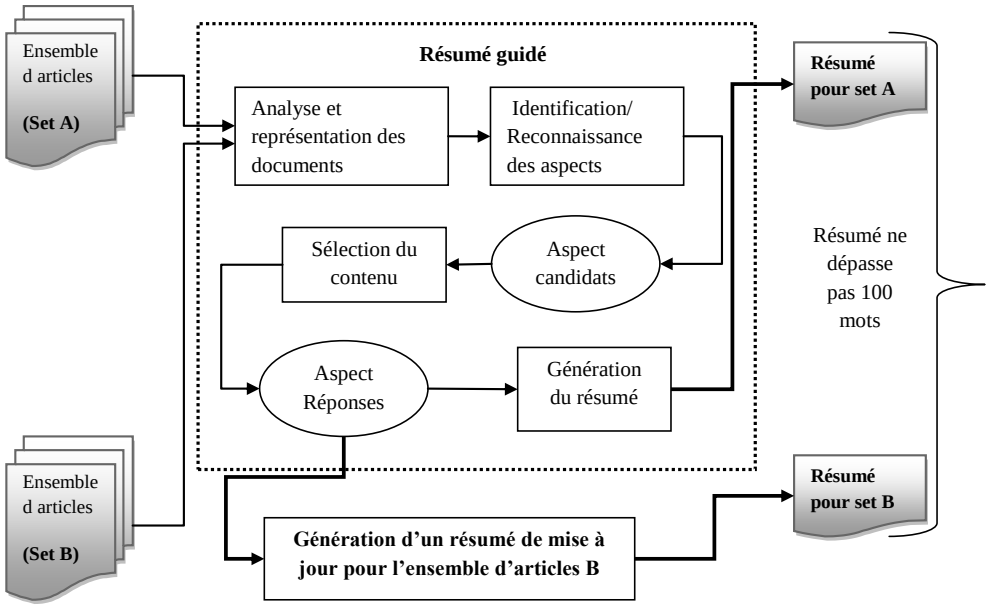

FIGURE 1 – Architecture générale d'un système de résumé guidé

Comme nous l'avons déjà mentionné dans les sections précédentes, un domaine d'application privilégié est le résumé guidé d'articles de presse circulant dans le web. Pour une catégorie d'articles donnée, les articles sont séparés en deux ensembles : *ensemble* A et *ensemble* B. D'abord, le système doit générer un résumé initial pour l'ensemble *d'articles* A, auquel le résumé doit répondre à tous les aspects prédéfinis. Puis, un résumé de mise à jour est généré pour l'ensemble *d'articles* B, en supposant que les documents de *l'ensemble* A ont été lus, et précédent chronologiquement les documents de *l'ensemble* B. Le résumé de mise à jour de *l'ensemble* B est également un résumé guidé, qui ne devrait pas contenir les informations déjà présentées dans le résumé de *l'ensemble* A. Chaque résumé doit être cohérent, concis, bien organisé, contenir des phrases grammaticalement correctes, et ne dépassant pas 100 mots.

Ainsi le résumé guidé peut être considéré comme un résumé multi-documents, de mise à jour, orienté par un ensemble d'aspects. Il regroupe trois types de résumés ce qui affirme que les types de résumé automatique de textes ne sont pas indépendants les uns des autres.

4 Systèmes de résumé automatique guidé

Dans la section 2, nous avons distingué trois grandes approches du résumé automatique en général les approches extractives, semi-extractives et abstractives. Dans cette section, nous présentons tout d'abord de façon succincte différents systèmes de résumé guidé se rapportant à ces trois grandes approches, en précisant les techniques utilisées pour chacun de ces systèmes. Ensuite, nous essayons de les comparer selon différents critères.

4.1 Systèmes basés sur l'approche extractive

Du *et al.* (2010) proposent deux méthodes nommées MRSP (Manifold Ranking with Sink Points) et TMSP (Topic guided Manifold Ranking with Sink Points). La méthode MRSP est dédiée au résumé de mise à jour, et vise à créer des résumés de haute qualité au niveau de la pertinence, l'importance, la nouveauté et la diversité de l'information. La méthode TMSP est une extension du MRSP qui intègre le modèle pLSA (Probablistic Latent Semantic Analysis) (Hofmann, 2013) avec un algorithme d'Espérance Maximum (EM) pour extraire les aspects. Les deux méthodes MRSP et TMSP sont basées sur l'approche Manifold Ranking (Zhou *et al.*, 2003).

Du *et al.* (2011) proposent une nouvelle méthode de classement de phrases DDRank (Decayed DivRank), une extension de DivRank (Mei *et al.*, 2010). Le modèle pLSA (Hofmann, 2013) est utilisé pour attribuer à chaque phrase un score mesurant sa probabilité d'appartenir à certains aspects. DDRank utilise les scores obtenus pour sélectionner les phrases du résumé. DDRank aborde la diversité, la pertinence et l'importance dans le classement de phrases d'une manière unifiée.

Zhang *et al.* (2011) développent deux systèmes de résumé : Polycom1 et Polycom2. Polycom1 est un système de base qui utilise une formule statistique pour attribuer des scores aux phrases sans prendre en considération les scores des aspects. Dans Polycom2 la reconnaissance de l'aspect au niveau de la phrase est considérée comme un problème de classification multilabels de textes auquel le modèle est construit en utilisant un nouveau type de caractéristiques (meta-phrase features), la technique de la décomposition binaire et le SVM (Vapnik, 1998; Joachims, 1999). Le modèle obtenu est ensuite utilisé pour prédire les phrases ayant des aspects et ces informations prédites sont alors utilisées pour calculer les scores des aspects dans les phrases. Polycom2 intègre les scores des aspects avec les scores obtenus par Polycom1 pour calculer les scores finaux des phrases.

Li *et al.* (2011a) proposent deux systèmes pour le résumé guidé : PKTUM1 et PKTUM2. Le premier combine linéairement les scores obtenus par l'algorithme Manifold Ranking avec des scores basés sur d'autres caractéristiques de surface pour pondérer et sélectionner les phrases pertinentes. Le deuxième est basé sur une variante de l'approche ILP (Integer Linear Programming) : T-ILP (Tolerated ILP) qui utilise Wikipedia comme domaine de connaissances pour améliorer la pondération des concepts. L'étape de sélection de phrases est précédée par un prétraitement et suivie par un post-traitement.

Barrera *et al.* (2011) proposent une méthode pour la génération des résumés guidés qui utilise un système Questions/Réponses SemQuest pour répondre aux aspects prédéfinis pour chaque catégorie.

Steinberger *et al.* (2010) proposent une nouvelle méthode pour le résumé guidé qui combine un système d'extraction d'information NEXUS (Tanev *et al.*, 2008) et la méthode LSA (Latent Semantic Analysis) (Gong & Liu, 2001; Steinberger & Jezek, 2004) pour capturer les informations pertinentes spécifiques aux aspects.

Varma *et al.* (2010) proposent une approche pour le résumé guidé qui se base sur un système d'extraction d'information et qui comporte principalement quatre étapes : la construction du domaine de connaissances, l'annotation de phrases pour identifier les informations spécifiques aux aspects, l'extraction des concepts pour calculer l'importance des phrases, et enfin la génération du résumé.

Conroy *et al.* (2011) développent le système CLASSY pour un résumé guidé multilingue qui combine des techniques statistiques avec le modèle naïf bayésien pour pondérer et sélectionner les informations pertinentes.

Zhang *et al.* (2012) améliorent le système qu'ils ont proposé en TAC'2011 (Zhang *et al.*, 2011), en ajoutant un modèle HMM (Hidden Markov Model) pour maximiser la cohérence des aspects trouvés.

Ng *et al.* (2012) proposent le système SWING (Guided Summarizer from WING "The Web Information Retrieval/Natural Language Processing") pour le résumé guidé qui se base sur les principes des systèmes de recherche d'information pour faire l'extraction des phrases pertinentes. L'idée principale est l'utilisation de la connaissance de catégorie (Category Knowledge), pour calculer l'importance spécifique des phrases par rapport à la catégorie (Category specific importance CSI). Le système Swing est classé parmi les meilleurs systèmes dans la campagne TAC'2011, et il pourra être appliqué à n'importe quelle catégorie, car les scores des caractéristiques utilisées se calculent à trois niveaux : le corpus, la catégorie et la thématique.

4.2 Systèmes basés sur l'approche semi-extractive

Li *et al.* (2011b) proposent une nouvelle approche par compression de phrases qui se divise en quatre étapes à savoir le clustering, le classement, la compression et la sélection de phrases. La première étape utilise un modèle LDA (Latent Dirichlet Allocation) pour identifier automatiquement les différents aspects et pour regrouper les phrases en ces aspects. La deuxième étape utilise une extension de l'algorithme LexRank (Erkan & Radev, 2004) pour classer les phrases dans chaque cluster. La troisième étape utilise un nouvel algorithme de compression de phrases pour améliorer la qualité linguistique des résumés. Enfin, la dernière étape utilise un Framework de programmation linéaire entière (ILP) pour sélectionner les phrases pertinentes et qui répondent aux aspects.

4.3 Systèmes basés sur l'approche abstractive

Peu de systèmes de résumé guidé existe, citons notamment le système ABSUM (Genest & Lapalme, 2012). Ce système implémente l'approche K-BABS (Knowledge-Based ABstractive Summarization) pour la génération automatique des résumés guidés par abstraction. L'architecture se constitue principalement de trois modules : le premier réalise une analyse fournissant une représentation intermédiaire des documents sources riche en information syntaxique et sémantique. Un deuxième module élabore un plan, appelé Task Blueprint qui définit manuellement les règles d'extraction d'information pour trouver les aspects candidats à partir de la représentation. Enfin,un troisième module fait une sélection de contenu sur les aspects candidats pour sélectionner les aspects réponses et générer ensuite le résumé par l'usage du logiciel SimpleNLG realizer (Gatt & Reiter, 2009).

4.4 Analyse comparative des systèmes de résumé automatique guidé

Dans cette sous-section, les systèmes brièvement présentés sont comparés selon divers critères, conduisant aux deux tableaux. Le premier tableau compare ces systèmes selon l'approche préconisée et les techniques mises en œuvre. Le second tableau compare ces systèmes selon les datasets qui ont été traités dans les publications les présentant, les mesures d'évaluation utilisées, les campagnes d'évaluation auxquelles ils ont participé, et enfin les résultats obtenus par ces systèmes dans ces campagnes.

TABLE 1 – comparaison des systèmes selon l'approche adoptée et les méthodes utilisées

Systèmes	Approche adoptée	Méthodes utilisées
Du *et al.* (2010)	extractive statistique	Algorithme Manifold Ranking - TMSP - MRSP - pLSA - l'algorithme Espérance-Maximisation
Steinberger *et al.* (2010)	extactive basée sur un système d'extraction d'événements	Le système NEXUS d'extraction - d'événements - LSA
Du *et al.* (2011)	extractive statistique	Algorithme DivRank - algorithme Decayed DivRank - pLSA
PolyCom (Zhang *et al.*, 2011)	extractive utilisant l'apprentissage automatique	SVM- Décomposition linéaire - (ISF) Inverted sentence frequency
PKTUM1/ PKTUM2 (Li *et al.*, 2011a)	extractive statistique	Algorithme Manifold Ranking ILP (Integer Linear Programming) - T-ILP (Tolerated-ILP)
Li *et al.* (2011b)	semi-extractive par compression de phrases utilisant l'apprentissage non-supervisé	Clustering - LDA (Latent Dirichlet Location) - l'algorithme LexRank - ILP (Integer Linear Programming)
CLASSY 2011 (Conroy *et al.*, 2011)	extractive hybride utilisant des techniques statistiques et l'apprentissage automatique	Naïf bayésien - caractéristiques linguistiques et statistiques
SWING (Ng *et al.*, 2012)	extractive hybride utilisant des techniques statistiques et l'apprentissage automatique	SVR - CRS (category relevance score) - CKLD (Category KL-divergence score) - INDF (Interpolated N-gram document frequency) - MMR
ABSUM (Genest & Lapalme, 2012)	abstractive (Template-based approach)	Analyse syntaxique et sémantique - Extraction de l'Information - NLG (Natural Language Generation)

D'après l'étude que nous avons menée sur les systèmes du résumé guidé, il est clair que *l'approche extractive* est celle qui a été retenue par la majorité des meilleurs systèmes. Rappelons que le principe de base de cette approche consiste à extraire des phrases pertinentes correspondantes aux aspects définis pour chaque catégorie définie pour le résumé guidé. Généralement, les premiers systèmes du résumé guidé exploitaient des méthodes statistiques, des méthodes d'apprentissage automatique, et des méthodes fondées sur la programmation linéaire avec une analyse de surface. Les résultats obtenus sont généralement intéressants, par exemple le système SWING (Ng *et al.*, 2012) fondé sur des méthodes statistiques est classé parmi les meilleurs systèmes dans TAC'2011. Le système CLASSY (Conroy *et al.*, 2011) basé à la fois sur des méthodes statistiques et sur un classificateur naïf bayésien est classé le premier en termes de *Overall Responsiviness*, et il atteint des scores intéressants par les autres méthodes d'évaluation ROUGE et Pyramide. En adoptant la même approche, d'autres systèmes améliorés ont été proposés, en introduisant des ressources sémantiques externes telles que Wordnet et Wikipedia, citons notamment le système PKTUM2 (Li *et al.*, 2011a). Basé sur une variante de la méthode ILP (Integer Linear Programming), celui-ci exploite Wikipidea comme

ressource sémantique, et il est classé premier par la méthode d'évaluation Pyramide et deuxième en termes de *Overall Responsiviness*.

D'autres auteurs (Li *et al.*, 2011b) ont opté pour *l'approche semi-extractive* en s'appuyant sur des techniques de compression de phrases, l'idée sous-jacente est de résoudre une des majeures problématiques du résumé par extraction, en éliminant les informations superflues et non essentielles contenues dans les phrases extraites. De plus elle permet d'établir un pont vers le résumé par abstraction. Le système proposé par Li *et al.* (2011b) a assuré des résultats intéressants.

L'approche abstractive constitue l'approche la plus récente, et la plus difficile à mettre en œuvre. Dans ce contexte, le système ABSUM proposé par (Genest & Lapalme, 2012) a atteint des résultats satisfaisants en termes de la qualité linguistique et du score Content Density. Cependant, les résumés générés par ABSUM se composent d'une moyenne de 21 mots contrairement aux 100 mots générés par les autres systèmes. Pour pallier cette limitation et améliorer le score du Overall Responsiveness du résumé, Genest & Lapalme (2012) ont introduit une approche hybride qui combine le système ABSUM et le système CLASSY. Les résultats obtenus montrent une amélioration significative en termes du Overall Responsiveness.

D'autres travaux (Barrera *et al.*, 2011) adoptent un système Questions-Réponses pour répondre aux aspects prédéfinis pour chaque catégorie. Une autre tendance consiste à aborder la problématique du résumé guidé comme plutôt un problème d'extraction d'information. Le système proposé par Varma *et al.* (2010), basé sur un système d'extraction d'information, a atteint le premier rang selon les méthodes ROUGE-2, ROUGE-SU4 et Pyramide, et il est classé le deuxième en termes du *Overall Responsiveness*.

5 Conclusion et perspectives

De l'analyse comparative précédente, on constate que les approches extractives sont actuellement dominantes dans le développement de systèmes automatiques de résumé guidé. Le principal avantage des méthodes extractives est qu'elles sont relativement simples à mettre en œuvre et qu'elles ne nécessitent pas une analyse en profondeur des textes, analyse assez complexe. L'inconvénient principal de ces méthodes est que les résumés produits manquent souvent de cohérence et de cohésion. En ce qui concerne les approches semi-abstractives, plus récentes, elles essaient de compenser les faiblesses des approches extractives sans pour autant les remettre en cause. Enfin les approches abstractives, bien que prometteuses, sont très difficiles à mettre en œuvre dans des systèmes automatiques de résumé. D'une façon générale, pour améliorer la qualité des résumés guidés obtenus par ces systèmes, il nous semble nécessaire de prendre en compte plus de sémantique, sémantique liée soit au domaine d'application du résumé guidé recherché, caractérisé par les catégories et leurs aspects, soit liée à la langue naturelle dans laquelle sont écrits les textes en entrée (source) et en sortie (si multilingue). Pour cela nous avons identifié deux grandes voies de recherche :

1. *Pour les approches extractives et semi-abstractives* : une amélioration majeure consisterait à adopter des représentations vectorielles enrichies des documents sources plus sémantiques, en s'appuyant notamment sur la désambiguïsation du sens (WSD), le calcul de similarité sémantique entre les mots. Cette approche est déjà utilisée dans le système développé par (Plaza *et al.*, 2010) pour résumer des documents biomédicaux, l'utilisation de la WSD permet d'améliorer les résultats obtenus en termes de performance. La construction des représentations lexicales distribuées pourrait également améliorer les systèmes du résumé. Ce type de représentations

peut se baser sur des techniques de réseaux de neurones dans le cadre de l'apprentissage profond. Ces techniques ont montré des résultats très intéressants en traitement automatique de la langue naturelle (Luong *et al.*, 2013; Zheng *et al.*, 2013), mais dans le résumé automatique très peu de travaux les utilisent jusqu'à présent, citons, cependant les travaux développés dans (Denil *et al.*, 2014).

2. *Pour l'approche abstractive* : ces approches utilisent déjà des représentations des documents sources plus sémantiques, il s'agirait principalement d'automatiser certaines tâches de traitement, actuellement réalisées de façon manuelle, ceci par la mise en œuvre de techniques récentes d'apprentissage machine. L'intérêt de l'apprentissage symbolique pour la réalisation de ces tâches est qu'il se situe au même niveau sémantique que celui pouvant être associé d'une part aux catégories et leurs aspects pouvant être liés à des connaissances externes comme des ontologies, et d'autre part à des connaissances linguistiques spécifiques. Ainsi, considérant le système (Genest & Lapalme, 2012), sa tâche de *blueprint* réalisant l'extraction des informations relatives aux aspects est actuellement réalisée de façon manuelle. Elle ne couvre pas toutes les catégories et elle nécessite assez de temps et d'efforts humains. Son automatisation, conduisant à l'automatisation de ce processus d'extraction d'information pouvant exploiter des ontologies, améliorerait de façon considérable ce système.

La production automatique de résumés guidés par abstraction reste un domaine jusqu'à présent peu exploré. Aux grand défis liés à la cohérence, la cohésion et la lisibilité des résumés générés, l'approche abstractive semble cependant la plus prometteuse. Très peu de systèmes ont été crées selon cette approche, citons FRUMP (DeJong, 1982), RIPTIDES (White *et al.*, 2001), et ABSUM (Genest & Lapalme, 2012). Tous ces systèmes combinent l'extraction d'information et les techniques de génération automatique de textes, mais reposent sur des tâches manuelles, notamment au niveau de l'extraction d'information.

Dans ce contexte, dans l'objectif de developper des systèmes automatiques de résumé guidé performants, notre recherche s'oriente plutôt dans la seconde voie, et aurait comme objectif la proposition d'une méthode abstractive pour le résumé guidé, composée de quatre étapes : (1) Analyse et la représentation des documents, (2) Extraction d'information, (3) Sélection de contenu, et (4) Génération du résumé. Cette méthode exploiterait une ontologie de domaine comme ressource sémantique externe pour guider le processus d'extraction d'information automatique, conformément à l'objectif du résumé guidé, l'utilisation des ontologies rend le contenu de résumé centré sur les besoins de l'utilisateur (Mohan *et al.*, 2016). L'étape d'*analyse et de la représentation des documents* se baserait sur des techniques linguistiques profondes qui exploiteraient la structure discursive de texte, notamment la théorie de la structure rhétorique (Mann & Thompson, 1988). L'étape d'*extraction d'information* reposerait sur l'utilisation d'une ontologie de domaine et sur une technique d'apprentissage automatique pour capturer les informations spécifiques aux aspects, comme le fait le système OntoILPER (Espinasse *et al.*, 2016) en utilisant la programmation logique inductive. L'étape de *sélection de contenu* exploiterait des ressources sémantiques externes (ontologies). Enfin, l'étape de *génération du résumé* serait assurée par le logiciel SimpleNLG Realizer (Gatt & Reiter, 2009), en prenant en considération l'importance des aspects et la relation entre eux. Une combinaison des approches abstractives et (semi)-extractives pourrait aussi être localement considérée.

TABLE 2 – Comparaison des principaux systèmes de résumé guidé

Systèmes	Datasets utilisés	Mesures d'évaluation	Campagne	Résultats
(Du *et al.*, 2010)	Apprentissage : TAC'2008/2009 test : TAC'2010	ROUGE-2 (R-2) ROUGE-SU4 (R-SU4) Pyramid Basic element (BE)	Les SRAT guidés participants à la compétition organisée par la campagne TAC'2010	Pour TMSP : Pyramid-A= 0,351 Rank (18), BE-A= 0.04529 Rank(21), R-2-A= 0.07623 Rank (21), R-SU4-A = 0.11042 Rank(21) Pour MRSP : Pyramid-B= 0,276 Rank(3), BE-B=0.04350 Rank(3), R-2-B=0.07251 Rank(4), R-SU4 - B = 0.11039 Rank(5)
(Du *et al.*, 2011)	TAC'2011	ROUGE-2 (R-2) ROUGE-SU4 (R-SU4) Pyramid Basic element (BE)	Les SRAT guidés participants à la compétition organisée par la campagne TAC'2011	Pour set A : Pyramid= 0.435 Rank (14) BE= 0.07099 Rank(13), R-A= 0.11324 Rank(11) R-SU4-A = 0.14901 Rank (9), Pour set B : Pyramid = 0.335 Rank(3), BE-B=0.05717 Rank(3), R-2 = 0.07992 Rank(14), R-SU4 = 0.12062 Rank(5)
PolyCom (aspect-integrated system) (Zhang *et al.*, 2011)	Apprentissage : TAC'2010 et un corpus crée manuellement à partir de DUC/TAC déjà passés Test : TAC'2011	ROUGE-2 (R-2) ROUGE-SU4 (R-SU4) Pyramid Basic element (BE) Linguistic Quality (LQ)	Les SRAT guidés participants à la compétition organisée par la campagne TAC'2011	Pour set A : R-2=0.12306 Rank (4), R-SU=0.15975 Rank(3), BE=0.07938 Rank(4), Pyramid=0,437 Rank(8), LQ=2,932 Rank(26) Pour set B : R-2=0.08643 Rank (4), R-SU=0.12803 Rank(2), BE=0.05437 Rank(9), Pyramid=0.3 Rank(17), LQ=2.795 Rank(25)
PKTUM1/ PKTUM2 (Li *et al.*, 2011a)	TAC'2011	ROUGE-2 (R-2) ROUGE-SU4 (R-SU4) Pyramid Basic element (BE) Linguistic Quality (LQ) Overall Responsiveness (OR)	Les SRAT guidés participants à la compétition organisée par la campagne TAC'2011	Pour PKTUM1 set A : R-2=0.102, R-SU=0.15975 Pyramid=0,418, LQ=3,136, OR= 2,977 (Rank13) Pour set B : R-2=, 0.0709, R-SU= 0.114, Pyramid=0.264, LQ= 3.023, OR= 2.432 (Rank15) Pour PKTUM2 set A : R-2= 0.115, R-SU= 0.150, Pyramid= 0.477 (Rank 1), LQ= 3.432, OR= 3.136 (Rank2) Pour set B : R-2= 0.0816, R-SU= 0.119, Pyramid= 0.313 LQ= 3.273, OR= 2.477 (Rank11)
(Li *et al.*, 2011b)	TAC'2010	ROUGE-1 ROUGE-2, ROUGE-L (R-L) ROUGE-SU4 ROUGE-W-1.2 (R-W-1.2)	K-means entity-aspect, greedy1, greedy2, KL-Div HIERSUM (Haghighi and Vanderwende, 2009),	R-1 = 0.32641, R-2= 0.06508,R-SU4= 0.10146, R-W-1.2 = 0.09998, R-L = 0.28610
CLASSY 2011 (Conroy *et al.*, 2011)	TAC'2011	Overall Responsiveness (OR) Linguistic Quality (LQ) Pyramid Content Density (CD) size	Les SRAT guidés participants à la compétition organisée par la campagne TAC'2011	Pyramid= 0.520, LQ= 3.39, OR= 3.20 Rank(1), size = 98.0%, CD= 0.0053
SWING (Ng *et al.*, 2012)	TAC'2011	ROUGE-2 ROUGE-SU4	Generic+ CRS, Generic +CKLD, CLASSY, PolyCom	R-2 = 0.13796, R-SU4 = 0.16808
ABSUM (Genest & Lapalme, 2012)	TAC'2011 Catégories : Attacks, Accidents and Natural disasters	Overall Responsiveness (OR) Linguistic Quality (LQ) Pyramid Content Density (CD) size	ABSUM, CLASSY 2011, ABSUM/CLASSY Hybrid, Extractive baseline, Abstractive baseline (Genest et Laplme, 2011), Human-written models	Pour ABSUM : Pyramid= 0.277 Rank (5), LQ= 3.67 (Rank 3), OR= 2.07 Rank(5), size = 22,6 % Rank(6), CD= 0.0119 Rank (1) Pour ABSUM/CLASSY hybrid : Pyramid= 0.600 Rank (1), LQ= 3.28 (Rank 4), OR= 3.31 Rank(2), size = 97.6% Rank(3), CD= 0.0061 Rank (2)

Références

ANDHALE N. & BEWOOR L. (2016). An overview of text summarization techniques. In *Computing Communication Control and automation (ICCUBEA), 2016 International Conference on*, p. 1–7 : IEEE.

BARALIS E., CAGLIERO L., MAHOTO N. A. & FIORI A. (2013). Graphsum : Discovering correlations among multiple terms for graph-based summarization. *Inf. Sci.*, **249**, 96–109.

BARRERA A., VERMA R. M. & VINCENT R. (2011). Semquest : University of houston's semantics-based question answering system. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*.

CARBONELL J. & GOLDSTEIN J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 335–336 : ACM.

CONROY J. M., SCHLESINGER J. D., KUBINA J., RANKEL P. A. & O'LEARY D. P. (2011). CLASSY 2011 at TAC : guided and multi-lingual summaries and evaluation metrics. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*.

CREMMINS E. (1993). Valuable and meaningful text summarization in thoughts, words, and deeds. In *Workshop on Summarizing Text for Intelligent Communication*.

CREMMINS E. (1996). *The Art of Abstracting*. Information Resources Press.

DEJONG G. (1982). An overview of the FRUMP system. In W. LEHNERT & M. RINGLE, Eds., *Strategies for Natural Language Processing*, p. 149–176. Lawrence Erlbaum.

DENIL M., DEMIRAJ A., KALCHBRENNER N., BLUNSOM P. & DE FREITAS N. (2014). Modelling, visualising and summarising documents with a single convolutional neural network. *CoRR*, **abs/1406.3830**.

DU P., YUAN J., LIN X., ZHANG J., GUO J. & CHENG X. (2011). Decayed divrank for guided summarization. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011*.

DU P., ZHANG J., GUO J. & CHENG X. (2010). TMSP : topic guided manifold ranking with sink points for guided summarization. In *Proceedings of the Third Text Analysis Conference, TAC 2010, Gaithersburg, Maryland, USA, November 15-16, 2010*.

ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, **22**, 457–479.

ERKAN G. & RADEV D. R. (2011). Lexrank : Graph-based lexical centrality as salience in text summarization. *CoRR*, **abs/1109.2128**.

ESPINASSE B., LIMA R. & FREITAS F. (2016). Extraction automatique d'entités et de relations par ontologies et programmation logique inductive. *Revue d'Intelligence Artificielle*, **30**(6), 637–674.

FATTAH M. A. (2014). A hybrid machine learning model for multi-document summarization. *Appl. Intell.*, **40**(4), 592–600.

GATT A. & REITER E. (2009). Simplenlg : A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, p. 90–93 : Association for Computational Linguistics.

GENEST P. & LAPALME G. (2012). Fully abstractive approach to guided summarization. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2 : Short Papers*, p. 354–358.

GENEST P., LAPALME G. & MONOD M. Y. (2009). HEXTAC : the creation of a manual extractive run. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*.

GENEST P.-E. & LAPALME G. (2011). Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, p. 64–73 : Association for Computational Linguistics.

GONG Y. & LIU X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR 2001 : Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, p. 19–25.

HOFMANN T. (2013). Probabilistic latent semantic analysis. *CoRR*, **abs/1301.6705**.

JIN F., HUANG M. & ZHU X. (2011). Guided structure-aware review summarization. *J. Comput. Sci. Technol.*, **26**(4), 676–684.

JOACHIMS T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, p. 200–209.

KNIGHT K. & MARCU D. (2000). Statistics-based summarization - step one : Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, July 30 - August 3, 2000, Austin, Texas, USA.*, p. 703–710.

KO Y. & SEO J. (2008). An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. *Pattern Recognition Letters*, **29**(9), 1366–1371.

LI H., HU Y., WAN X., XIAO J. & LI Z. (2011a). PKUTM participation at TAC 2011 summarization track. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*.

LI P., WANG Y., GAO W. & JIANG J. (2011b). Generating aspect-oriented multi-document summarization with event-aspect model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 1137–1146.

LI S., WANG W. & ZHANG Y. (2009). TAC 2009 update summarization of ICL. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In S. S. MARIE-FRANCINE MOENS, Ed., *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LLORET E. & PALOMAR M. (2012). Text summarisation in progress : a literature review. *Artif. Intell. Rev.*, **37**(1), 1–41.

LUHN H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, **2**(2), 159–165.

LUONG T., SOCHER R. & MANNING C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 104–113.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, **8**(3), 243–281.

MEI Q., GUO J. & RADEV D. (2010). Divrank : the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 1009–1018 : Acm.

MIHALCEA R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, July 21-26, 2004 - Poster and Demonstration*.

MOHAN M. J., SUNITHA C., GANESH A. & JAYA A. (2016). A study on ontology based abstractive summarization. *Procedia Computer Science*, **87**, 32–37.

NALLAPATI R., ZHAI F. & ZHOU B. (2017). Summarunner : A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, p. 3075–3081.

NALLAPATI R., ZHOU B., DOS SANTOS C. N., GÜLÇEHRE Ç. & XIANG B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, p. 280–290.

NENKOVA A. & MCKEOWN K. (2012). A survey of text summarization techniques. In *Mining Text Data*, p. 43–76.

NENKOVA A. & PASSONNEAU R. J. (2004). Evaluating content selection in summarization : The pyramid method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, p. 145–152.

NG J., BYSANI P., LIN Z., KAN M. & TAN C. L. (2012). Exploiting category-specific information for multi-document summarization. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, 8-15 December 2012, Mumbai, India*, p. 2093–2108.

OLIVEIRA H., FERREIRA R., LIMA R., LINS R. D., FREITAS F., RISS M. & SIMSKE S. J. (2016a). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Syst. Appl.*, **65**, 68–86.

OLIVEIRA H., LIMA R., LINS R. D., FREITAS F., RISS M. & SIMSKE S. J. (2016b). Assessing concept weighting in integer linear programming based single-document summarization. In *Proceedings of the 2016 ACM Symposium on Document Engineering, DocEng 2016, Vienna, Austria, September 13 - 16, 2016*, p. 205–208.

PLAZA L., STEVENSON M. & DÍAZ A. (2010). Improving summarization of biomedical documents using word sense disambiguation. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, p. 55–63 : Association for Computational Linguistics.

RUSH A. M., CHOPRA S. & WESTON J. (2015). A neural attention model for abstractive sentence summarization. *CoRR*, **abs/1509.00685**.

SPARCK J. K. (1998). Automatic summarising : factors and directions. *CoRR*, **cmp-lg/9805011**.

STEINBERGER J. & JEZEK K. (2004). Text summarization and singular value decomposition. In *Advances in Information Systems, Third International Conference, ADVIS 2004, Izmir, Turkey, October 20-22, 2004, Proceedings*, p. 245–254.

STEINBERGER J., TANEV H., KABADJOV M. A. & STEINBERGER R. (2010). Jrc's participation in the guided summarization task at TAC 2010. In *Proceedings of the Third Text Analysis Conference, TAC 2010, Gaithersburg, Maryland, USA, November 15-16, 2010*.

TANEV H., PISKORSKI J. & ATKINSON M. (2008). Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems, 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008, London, UK, Proceedings*, p. 207–218.

TORRES-MORENO J.-M. (2014). *Automatic text summarization*. John Wiley & Sons.

TZOURIDIS E., NASIR J. & BREFELD U. (2014). Learning to summarise related sentences. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 1636–1647.

VAPNIK V. (1998). *Statistical learning theory*. Wiley.

VARMA V., BYSANI P., B K. R., REDDY V. B., KOVELAMUDI S., VADDEPALLY S. R., NANDURI R., KUMAR N. K., GSK S. & PINGALI P. (2010). IIIT hyderabad in guided summarization and knowledge base population. In *Proceedings of the Third Text Analysis Conference, TAC 2010, Gaithersburg, Maryland, USA, November 15-16, 2010*.

WHITE M., KORELSKY T., CARDIE C., NG V., PIERCE D. & WAGSTAFF K. (2001). Multidocument summarization via information extraction. In *Proceedings of the first international conference on Human language technology research*, p. 1–7 : Association for Computational Linguistics.

YANG L., CAI X., ZHANG Y. & SHI P. (2014). Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization. *Inf. Sci.*, **260**, 37–50.

ZAJIC D. M., DORR B. J. & LIN J. (2008). Single-document and multi-document summarization techniques for email threads using sentence compression. *Information Processing & Management*, **44**(4), 1600–1610.

ZHANG R., LI W. & GAO D. (2012). Generating coherent summaries with textual aspects. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.*

ZHANG R., YOU O. & LI W. (2011). Guided summarization with aspect recognition. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*.

ZHENG X., CHEN H. & XU T. (2013). Deep learning for chinese word segmentation and pos tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 647–657.

ZHOU D., WESTON J., GRETTON A., BOUSQUET O. & SCHÖLKOPF B. (2003). Ranking on data manifolds. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS]*, p. 169–176.