Une nouvelle métrique d'évaluation pour les résumés automatiques de texte par extraction

Ahmed Chaouki Lokbani*
Mohamed Amine Boudia**
Reda Mohamed Hamou***

*GeCode Laboratory
Department of Computer Science Tahar Moulay University of Saida Algeria
ahchlokbani@yahoo.fr

** GeCode Laboratory
Department of Computer Science Tahar Moulay University of Saida Algeria
mamiamounti@yahoo.fr

*** GeCode Laboratory
Department of Computer Science Tahar Moulay University of Saida Algeria
hamoureda@yahoo.fr

Résumé. Dans cet article, nous proposons une métrique d'évaluation pour les résumés automatiques de textes en occurrence là une adaptation de la F-Mesure qui engendre une hybridation de méthode d'évaluation d'un résumé automatique au même temps : Extrinsèque et Intrinsèque. Notre travail suit la méthodologie suivante : on commence par l'étude de faisabilité d'adaptation de la F-Mesure pour l'évaluation des résumés automatiques, ensuite en on dé-finira la façon de calculer la F-Mesure pour un résumé candidat. Les textes sont représentés par un vecteur de terme qui peut être soit une phrase soit un mot (avec une pondération binaire ou par occurrence). Afin de déterminer l'exactitude de l'évaluation F.-Mesure pour les résumés automatiques par ex-traction, on calcule de corrélation avec l'évaluation ROUGE.

1 Introduction

Un résumé est un texte est le faite de rééditer le texte en taille plus réduite. Sous contrainte gardée la sémantique d'un document autrement dit minimise l'entropie sémantique. Le but de cette opération est d'aider le lecteur à repérer les informations intéressantes pour lui sans être obligé à lire entièrement le document.

« On ne peut pas imaginer dans notre vie quotidienne, une journée sans résumé », souligne Inderjeet Mani (Mani et al., 2002). Les titres des journaux, le premier paragraphe d'un article de journal, les bulletins d'informations, les prévisions météo, les tables des résultats des compétitions sportives et les catalogues des bibliothèques sont tous des résumés. Même dans la recherche, les auteurs d'article scientifique doivent accompagner leurs articles scientifiques par des résumés écrits par eux-mêmes.

Le volume d'information textuelle électronique ne cesse d'augmenter, rend l'accès à l'information une tâche difficile. La production d'un résumé peut faciliter l'accès à l'information

mais elle est aussi une tâche complexe car elle nécessite des connaissances linguistiques (Boudin et Morin, 2013).

Ainsi la complexité de résumé automatique s'avère dans la partie d'évaluation la qualité de résumé produit ; la communauté de traitement Automatique de Langage Naturel n'a pour l'instant une solution précise à ce problème et propose que des solutions partielles. Et en réalité, il n'existe pas de résumé « idéal ».

Dans la littérature en trouve deux façons d'évaluation de résumé automatiques :

— extrinsèque,
— Intrinsèque,

Notre travail consisté à l'adaptation de la mesure F.-Mesure pour l'évaluation de qualité d'un résumé automatique, cette adaptation engendre une hybridation des méthode d'évaluation (Extrinsèque et Intrinsèque), globalement notre travail répond aux questions suivantes :

— est-ce qu'on peut appliquer l'évaluation F.-mesure pour les résumés automatiques.
— Comment on va adapter la F-Mesure pour l'appliquer sur les résumés automatiques ?
— Est-ce que résultat d'évaluation de la F-Mesure reflète vraiment la qualité de résumé automatique ?

2 Notre approche

Le résumé automatique par extraction peut être vu comme une classification bi-classe (« unités textuelles à garder » et « unités textuelles à supprimer») : les techniques de résumé automatique par extraction sont basées sur le choix des unités textuelles de texte (phrase par défaut) selon leur importance ou selon leur corrélation avec la thématique de texte complet ou suivante une autre mesure.
Afin de valider les résultats de classification supervisée, la mesure F-Mesure est recommandée parce qu'elle se base sur deux indices : Rappel et précision.

Le rappel (en classification) est défini par le nombre de documents pertinents trouvés (par le classifieur) sur le nombre de documents pertinents que possède la base de données. Cela signifie que lorsque l'utilisateur interroge la base il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le système documents requête est importante alors le taux de rappel est élevée. À l'inverse si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel.

La précision (en classification) est le nombre de documents pertinents trouvés (par le classifieur) sur le nombre de documents total proposés par le classifieur pour une requête donnée. Le principe est le suivant : quand un utilisateur interroge une base de données, il souhaite que les documents proposés en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme "précis".

Notre idée est d'adapter la mesure d'évaluation F.mesure au résumé automatique, pour cela nous avons commencé par faire la comparaison suivante (voir 1) :

	La Classification	**Résumé automatique par extraction**
Entrée INPUT	Des documents	Un texte ou plusieurs textes
Résultat OUTPUT	Des classes telles que le nombre de classes sont connues au préalable	Un texte (un résumé), d'une autre vu deux classes : • unité textuelle a gardé • unité textuelle a supprimé
Mode Evaluation	Automatique (la classe réelle des documents est connue)	Semi-automatique (Par des résumés référencés produits par humain aux autres systèmes de R. A)

TAB. 1 – *La classification vs Résumé automatique*

En se basant sur cette comparaison confirme la possibilité d'adaptation de F-Mesure pour évaluer les résumés automatiques. Le fond de problème de l'adaptation est alors de proposer une matrice de confusion ; à partir de cette matrice de confusion on peut calculer : le rappel et la précision donc le F-Mesure.

Notre méthode est une hybridation entre les deux méthodes d'évaluation : intrinsèque et extrinsèque.

On procède à comparer le résumé candidat et le texte complet (à résumé) afin d'identifier les unités textuelles qui ont été gardées et celle qui ont été supprimée, ensuite on refait la même opération entre le résumé référence et le texte complet (à résumé), et enfin on effectue une comparaison entre les deux résumé (candidat et référence) afin d'obtenir la matrice de confusion suivante (voir 1) :

Matrice de confusion Classification		Classe éstimé C1	Classe éstimé C2
Classe réelle	C1	A	B
	C2	C	D

⇒

Matrice de confusion Résumé Automatique		Résumé candidat Ut à G	Résumé candidat Ut à S
Résumé référence	Ut à G	X	Y
	Ut à S	Z	W

Ut à G : unités textuelles à garder

Ut à S : unités textuelles à supprimés

FIG. 1 – *Passage de la matrice de confusion de classification vers résumé automatique*

Classification	Résumé Automatique
A : Vrai Positive (VP) : les documents qui appartienne réellement à C1 et que le modèle les a classé en C1.	X : Vrai Positive (VP) : les unités textuelle gardés dans le résumé référence, et aussi gardés dans le résumé candidat.
B : Faux Négative (FN) : les documents qui appartienne réellement à C1 et que le modèle les a classé en C2.	Y : Faux Négative (FN) : les unités textuelle gardés dans le résumé référence, mais supprimés dans le résumé candidat.
C : Faux Positive (FP) : les documents qui appartienne réellement à C2 et que le modèle les a classé en C1.	Z : Faux Positive (FP) : les unités textuelle supprimés dans le résumé référence, mais gardés dans le résumé candidat.
D : Vrai Négative (VN) : les documents qui appartienne réellement à C2 et que le modèle les a classé en C2.	W : Vrai Négative (VN) : les unités textuelle supprimé dans le résumé référence, et aussi supprimé dans le résumé candidat.

TAB. 2 – *Interprétation de la matrice de confusion de classification vs résumé automatique*

La précision de la classe « Ut à G » est défini par le nombre des unités textuelles gardées dans le résumé candidat et référence (commun), divisé sur le nombre de ces unités textuelles gardées par le résumé candidat ; en parallèle en calcule la précision de la classe « Ut à S » de la même façon c'est-à-dire : le nombre des unités textuelles supprimé dans le résumé candidat et référence (commun), divisé sur le nombre de ces unités textuelles supprimé par le résumé candidat.

$$Rappel_{UtG} = \frac{X}{X+Y} \tag{1}$$

$$Rappel_{UtS} = \frac{W}{W+Z} \tag{2}$$

$$Prcision_{UtG} = \frac{X}{X+Z} \tag{3}$$

$$Prcision_{UtS} = \frac{W}{W+Y} \tag{4}$$

Puisque le résumé automatique par extraction est une classification bi-classes donc :

$$Rappel = \frac{Rappel_{UtG} + Rappel_{UtS}}{2} \tag{5}$$

$$Prcision = \frac{Prcision_{UtG} + Prcision_{UtS}}{2} \tag{6}$$

Enfin en combinant la précision et le rappel et leur pondération afin de calculer la F-Mesure

$$F - mesure = \frac{2 \times Prcision \times Rappel}{Prcision + Rappel} \tag{7}$$

Pour procéder à l'évacuation de résumé automatique par la F-Mesure on doit passer par la phase de représentation de texte : dans notre cas où peut utiliser trois représentations :

—représentation en sac de mots (binaire).

—représentation en sac de mots (occurrence).

—représentation en sac de phrase.

3 Expérimentation

Afin d'expérimenter la mesure proposée nous avant fixé l'unité textuelle à un mot (sac de mot) puis a une phrase (sac de phrase) ; pour chaque choix nous avons comparé les résultats obtenus avec ceux de la mesure ROUGE-SM(2).

La construction de matrice de confusion dépend sur la détermination d'unité textuelle, pour l'expérimentation trois modèles de construction de matrice de confusion ont été proposés ci-dessus.

3.1 Corpus Utilisé

On a utilisé comme corpus le texte « mars » en langue française qui contient un titre et 11 phrases, après les processus de prétraitement et de vectorisation, nous obtenons 208 termes.

Les résumés candidat : les résumés obtenus par notre résumeur basé sur les araignées sociales pour les résumés automatiques de textes. (choisi d'une manière aléatoire) (Boudia et al., 2015).

Les résumés référence : on a pris trois résumés référence produit successivement par le résumeur REG, CORTEX et le troisième par un expert humain.

3.2 Validation

Afin de déterminer l'exactitude, la robustesse et la pertinence de la métrique F-Mesure vis-à-vis de l'évaluation des résumés automatiques par extraction, un calcul de corrélation avec métrique d'évaluation existante et valide s'impose systématiquement. Nous avons calculé la corrélation avec la métrique ROUGE-SU(2) proposé par Lin en 2004 (Lin, 2004) qui est reconnue et largement utilisée lors de conférence DUC .

Recall-Oriented Understudy for Gisting Evaluation où en abréviation ROUGE a été introduite par Lin en 2004 (Lin et al., 2006), où il propose d'évaluer les résumés candidats d'une manière semi-automatique en mesurent les similarités entre un résumé candidat et un ou plusieurs résumés de références.

ROUGE-(N)

$$Rouge(n) = \frac{\sum_{s \in R_{ref}} \sum_{s \in R_{can}} Co-Occurence(R_{ref}, R_{ref})}{NbrNGram(N)_{ref}} \tag{8}$$

ROUGE-SU(M) : Adaptation de ROUGE-2 utilisant des bigrammes à trous (skip units, (SU)) de taille maximum M et comptabilisant les unigrammes.

Nous avons utilisé le coefficient de corrélation linéaire de Bravais-Pearson qui permet de détecter l'existence ou l'absence et de mesurer la force d'une relation linéaire entre deux items qui doit être quantitatifs et continus Louis et Nenkova (2009).

Le calcule de la covariance est pré-requis. Rappelons que la covariance représente la moyenne du produit des écarts à la moyenne.

$$Cov(X,Y) = \frac{1}{N}\sum_{i=0}^{N}(X_i - \overline{X}).(Y - \overline{Y}) \tag{9}$$

Le coefficient de corrélation linéaire de deux caractères X et Y est égal à la covariance de X et Y divisée par le produit des écarts-types de X et Y.

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x.\sigma_y} \tag{10}$$

Ci dessous un tableau (Voir Tableau 3) qui explique l'interprétation de la valeur de corrélation linéaire de Bravais-Pearson.

Corrélation	**Négative**	**Positive**
Faible	De -0,5 à 0.0	De 0,0 à 0.5
Forte	De -1,0 à -0.5	De 0,5 à 1.0

TAB. 3 – *Interprétation de la valeur de corrélation linéaire de Bravais-Pearson*

3.3 Résultats et interprétation

Après expérimentations nous avons regroupé les résultats dans les tableaux ci-dessous. (Voir Tableau 4 jusqu'à Tableau **??**)

1/ Premier résumé candidat

Résumé candidat 1 VS Cortex (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE- SU(2)
	Résumé Candidat					
Résumé Référence	5	1	0,666	0,6875	0,6769	0,69
	3	3				
Résumé candidat 1 VS Cortex (référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	69	14	0,5328	0,5509	0,5417	0,69
	49	15				
Résumé candidat 1 VS Cortex (référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	90	14	0,6209	0,6625	0,641	0,69
	53	32				

TAB. 4 – *Résultats d'évaluation du premier résumé candidat (le résumé référence cortex) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

Résumé candidat 1 VS REG (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	5	1	0,666	0,6875	0,6769	0,75
	3	3				
Résumé candidat 1 VS REG(référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	87	10	0,6384	0,6962	0,66609	0,75
	31	19				
Résumé candidat 1 VS REG(référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	124	11	0,7166	0,7753	0,7481	0,75
	33	35				

TAB. 5 – *Résultat d'évaluation du premier résumé candidat(le résumé référence REG) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

Résumé candidat 1 VS Humain (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	3	0	0.7222	0.6875	0.7044	0,66
	5	4				
Résumé candidat 1 VS Humain (référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	85	10	0,63006	0,6877	0,6576	0,66
	33	19				
Résumé candidat 1 VS Humain (référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	124	14	0,6955	0,7434	0,7167	0,66
	34	33				

TAB. 6 – *Résultats d'évaluation du premier résumé candidat (le résumé référence : l'humain) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

2/ deuxième résumé candidat

Résumé candidat 2 VS Cortex (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	4	2	0,6666	0,6666	0,6666	0,7
	2	4				
Résumé candidat 2 VS Cortex (référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	58	25	0,5681	0,5726	0,5703	0,7
	36	28				
Résumé candidat 2 VS Cortex (référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	74	26	0,6952	0,6965	0,6959	0,7
	36	67				

TAB. 7 – *Résultats d'évaluation du deuxième résumé candidat (le résumé référence cortex) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

Résumé candidat 2 VS REG (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	3	3	0,5	0,5	0,5	0,6
	3	3				
Résumé candidat 2 VS REG(référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	63	34	0,5147	0,5143	0,5145	0,6
	31	19				
Résumé candidat 2 VS REG(référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	84	46	0,6142	0,6089	0,6115	0,6
	33	46				

TAB. 8 – *Résultats d'évaluation du deuxième résumé candidat(le résumé référence REG) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

Résumé candidat 2 VS Humain (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	3	0	0,8333	0,75	0,7894	0,69
	3	6				
Résumé candidat 2 VS Humain (référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	69	26	0,6217	0,6277	0,6222	0,69
	25	27				
Résumé candidat 2 VS Humain (référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	96	36	0,7093	0,7010	0,7051	0,69
	25	56				

TAB. 9 – *Résultats d'évaluation du deuxième résumé candidat (le résumé référence : l'humain) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

3/ Troisième résumé candidat

Résumé candidat 3 VS Cortex (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	4	2	0,5833	0,5857	0,5842	0,61
	3	3				
Résumé candidat 3 VS Cortex (référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	60	23	0,4786	0,4725	0,4755	0,61
	49	15				
Résumé candidat 3 VS Cortex (référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	78	26	0,5839	0,5945	0,5891	0,61
	53	38				

TAB. 10 – *Résultats d'évaluation du troisième résumé candidat (le résumé référence cortex) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

Résumé candidat 3 VS REG (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	4	2	0,5833	0,5857	0,5845	0,69
	3	3				
Résumé candidat 3 VS REG(référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	78	19	0,5920	0,6077	0,5998	0,69
	31	19				
Résumé candidat 3 VS REG(référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	110	24	0,6812	0,6941	0,6876	0,69
	33	39				

TAB. 11 – *Résultats d'évaluation du troisième résumé candidat(le résumé référence REG) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

Résumé candidat 3 VS Humain (référence) Modèle 1 : unité textuelle = Phrase						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	3	0	0,7777	0,7142	0,7446	0,68
	4	5				
Résumé candidat 3 VS Humain(référence) Modèle 2 : unité textuelle = Mot (binaire)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	76	19	0,5826	0,5986	0,5905	0,68
	33	19				
Résumé candidat 3 VS Humain(référence) Modèle 3 : unité textuelle = Mot (par occurrence)						
Matrice de confusion			Rappel	Précision	F-Mesure	ROUGE-SU(2)
	Résumé Candidat					
Résumé Référence	111	27	0,6660	0,675	0,6705	0,68
	34	38				

TAB. 12 – *Résultats d'évaluation du troisième résumé candidat (le résumé référence : l'humain) en utilisant les trois modèles de représentation textuelle, et la comparaison avec ROUGE-SU(2)*

Voici ci dessus le calcule de corrélation entre F-mesure et ROUGE (Voir Figure 2) :

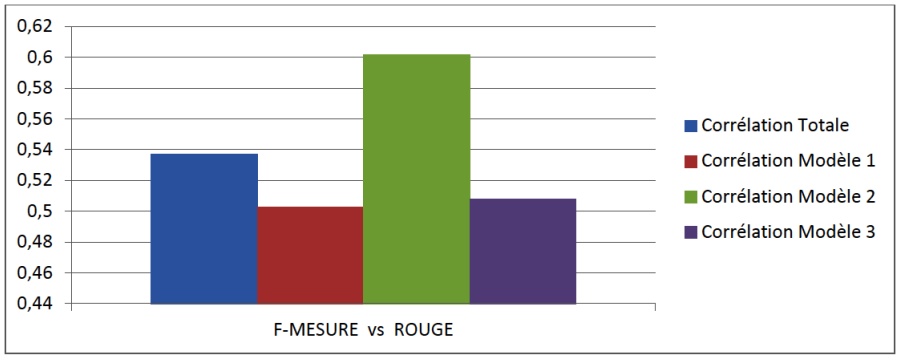

FIG. 2 – *La corrélation entre F-mesure et ROUGE-SU (2) avec les différents modèles de représentation textuelle*

La métrique ROUGE est une méthode d'évaluation semi-automatique intrinsèque : elle se base sur les similarités entre un résumé candidat et un ou plusieurs résumés références. Son principale inconvenient est qu'elle ne revient pas au texte original et se base uniquement sur les résumés référence.

La F-Mesure est l'une des métriques les plus robustes utilisées pour l'évaluation des classifications ; la F-Mesure est une combinaison de Rappel et Précession. Pour notre adaptation nous ajoutons à la force de F-Mesure le fait de procèder à une évaluation extrinsèque au début, et nous enchaînons avec une évaluation intrinsèque : donc une évaluation hybride par une projection sur la théorie de l'évaluation des résumés automatiques par extraction.

En effet, la première étape de génération de matrice de confusions consiste à comparer le résumé candidat et les résumés références avec le texte original afin de détecter les unités textuelles du texte original qui ont été gardées dans le résumé candidat ainsi que les résumés références, puis les unités textuelles du texte original qui ont été supprimées dans le résumé candidat ainsi que les résumés références ; cela est l'évaluation extrinsèque par définition : réalisée par une machine où le résumé candidat est évaluer par rapport au texte orignal dans le contexte d'une tâche spécifique, dans notre cas la tâche est une classification supervisé des unité textuelle en deux classes : "les unités textuelles gardées" et "les unités textuelles supprimées". La deuxième étape de génération de matrice de confusion est d'effectuer une seconde comparaison entre les résultats de comparaison des résumés références au texte original résultante de la première étape, avec les résultats de comparaison du résumé candidat avec le texte original également résultante de la première étape, cela est l'évaluation intrinsèque où un ou plusieurs résumés candidats sont comparés à un ou plusieurs résumés de référence.

La précision indique la pureté du résumé candidat tandis que le rappel interprète la ressemblance entre le résumé candidat et le résumé référence.

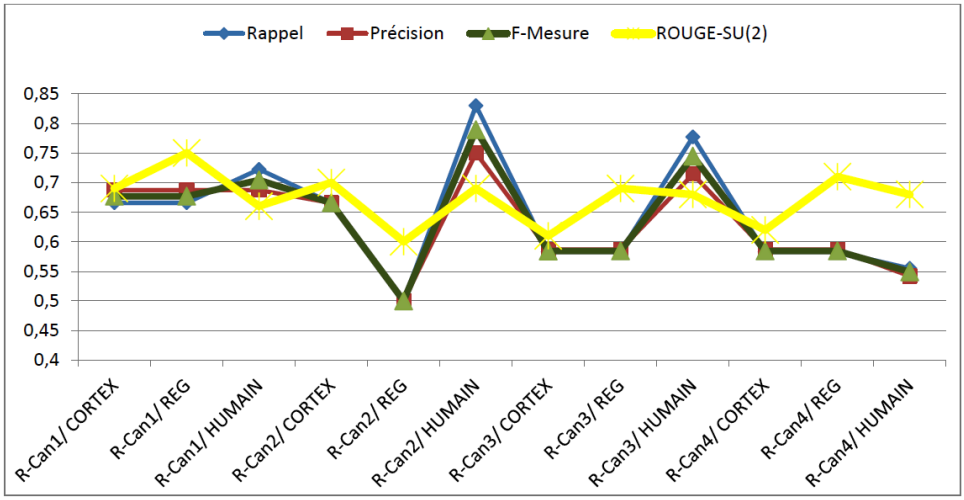

FIG. 3 – *Rappel, Précision et F-Mesure avec Modèle 1(UT=Phrase) Vs ROUGE-SU(2)*

À la lecture de ce graphe (Voir Figure 3) nous constatons : premièrement, que le rappel est souvent plus grand que la précision et nous remarquons que les résultats obtenus par la F-Mesure calculés en choisissant "sac de phrases" comme méthode du choix de terme , ne correspond pas au ROUGE-SU (2), et nous constatons aussi que les résultats d'évaluation ne sont pas stables par rapport à la métrique ROUGE-SU(2) : parfois la F-Mesure est élevée par rapport ROUGE-SU(2) et vise versa. Ce qui explique que la corrélation du premier modèle de choix d'unité textuelle en l'occurrence « une phrase » est le plus petit parmi les trois corrélations.

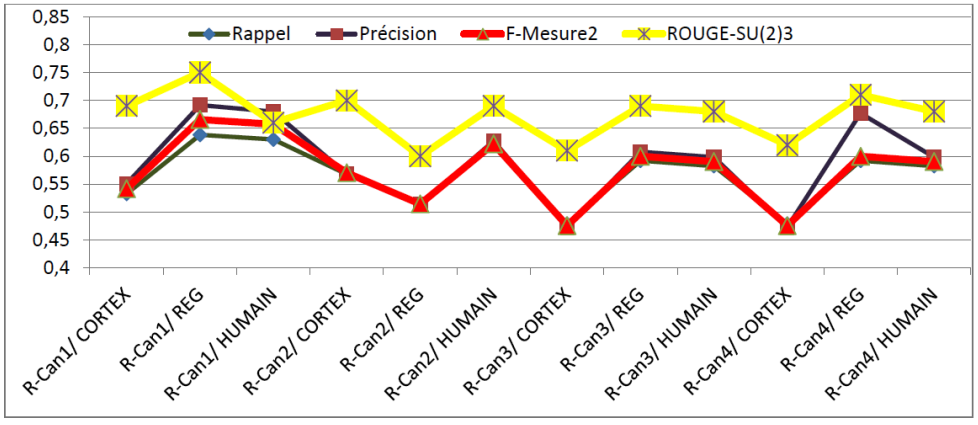

FIG. 4 – *Rappel, Précision et F-Mesure avec Modèle 2 (UT=Mot-Traitement=binaire) Vs ROUGE-SU(2)*

Contrairement au modèle de choix des unités textuelles précédent, nous voyons clairement que le choix de terme du deuxième modèle « unité textuelle = un mot traitement = binaire» a

nettement amélioré les résultats obtenus par la F-Mesure par rapport à la métrique ROUGE, dans le graphe (Voir Figure 4) nous constatons que les deux courbes sont presque parallèles, ce qui est démontré par l'indice de corrélation qui est le plus élevé (0,60). Nous constatons que même le rappel et la précision sont corrélés (selon le graphe) avec l'indice ROUGE-SU(2) et nous retenons de cela aussi que la précision est plus élevée que le rappel pour ce modèle, ce qui veux dire qu'il y a plus de silence que de bruit.

Dans le graphe suivant (Voir Figure 5) la différence entre ce modèle de représentation textuelle basé sur le sac de mot et la pondération basée sur l'occurrence d'appariation et les modèles précédents, réside dans le fait qu'en changeant la façon d'incrémentation de compteur (VP, FP, FN, VN) du « binaire » à « par occurrence » a totalement changé les résultats obtenus par F-Mesure par rapport à la métrique ROUGE-SU(2) , même si les valeurs se sont rapprochées, les courbes ne sont pas parallèles, ce qui est montré par l'indice de corrélation classé en deuxième position. Nous pouvons voir clairement que la précision est très élevée et que le rappel est inférieur à la F-Mesure.

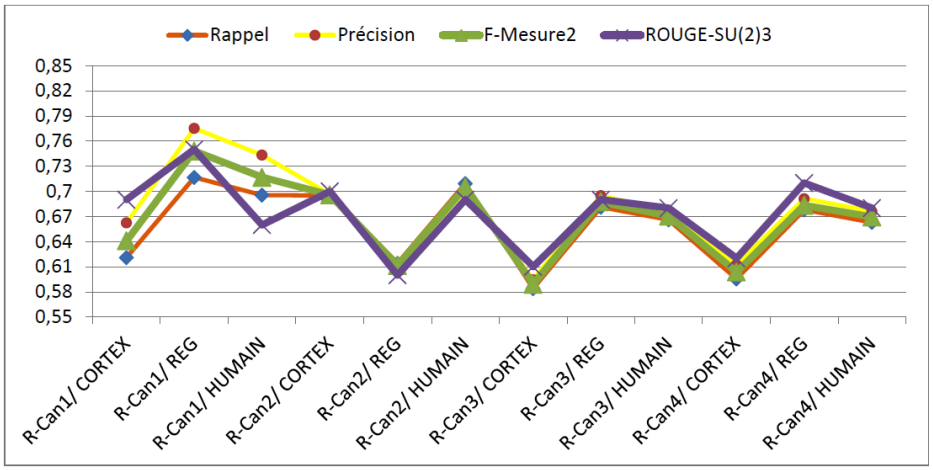

FIG. 5 – *Rappel, Précision et F-Mesure avec Modèle 3 (UT=Mot-Traitement=occurrence) Vs ROUGE-SU(2)*

4 Les limites de la mesure proposée

ROUGE est une métrique d'évaluation semi-automatique intrinsèque qui se base sur le nombre de co-occurrences entre un résumé candidat et un ou plusieurs résumés références divisé par la taille de ces derniers. Sa faiblesse est qu'elle ne se base que sur les résumés références et néglige le texte orignal.

Les valeurs données par ROUGE pour un résumé, à un taux de réduction négligeable, est élevé. Cette valeur élevée s'explique par l'augmentation du nombre de co-occurrences des termes entre le résumé candidat et les résumés références. La F-Mesure est l'une des métriques les plus robustes utilisées pour l'évaluation de la classification ; la F-Mesure est une combinaison de Rappel et de Précision. Pour l'adaptation, on ajoute à la force de F-Mesure,

une évaluation extrinsèque au début, et on enchaîne avec une évaluation intrinsèque : donc une évaluation hybride. Pour un résumé automatique à taux de réduction réduit, la F-Mesure donne des évaluations meilleures que celles de ROUGE car elle prend en considération l'absence des termes. Mais contrairement à ROUGE l'évaluation du résumé candidat à taux de réduction élevé peut être faussée puisque la valeur de faux négative sera maximale ce qui donnera de bonnes évaluations pour des résumés généralement médiocres notant que le taux de réduction élevée entraine une augmentation d'entropie d'information.

5 Conclusion

Dans ce travail, nous avons présenté un nouvel outil de validation des résumés automatiques des textes par extraction, en l'occurrence la mesure F-Mesure. En première intention, nous avons voulu démontrer la faisabilité d'adapter la F-Mesure. La deuxième intention, est de réaliser une hybridation des méthodes d'évaluation connues pour les résumés automatiques, en occurrence : intrinsèque et extrinsèque. Vu les résultats obtenus, notre adaptation peut contribuer à résoudre une des problématiques majeures des résumés automatiques : l'évaluation et la validation des résumeurs.

Références

Boudia, M. A., R. M. Hamou, A. Amine, M. E. Rahmani, et A. Rahmani (2015). A new multi-layered approach for automatic text summaries mono-document based on social spiders. In *IFIP International Conference on Computer Science and its Applications_x000D_*, pp. 193–204. Springer.

Boudin, F. et E. Morin (2013). Keyphrase extraction for n-best reranking in multi-sentence compression. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Lin, C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Lin, C.-Y., G. Cao, J. Gao, et J.-Y. Nie (2006). An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 463–470. Association for Computational Linguistics.

Louis, A. et A. Nenkova (2009). Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1-Volume 1*, pp. 306–314. Association for Computational Linguistics.

Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, et B. Sundheim (2002). Summac : a text summarization evaluation. *Natural Language Engineering 8*(1), 43–68.

Summary

In this paper, we propose a new metric of evaluation for automatic summaries of texts in this case the adaptation of the F-measure that generates a hybrid method of evaluating an automatic summary at the same time: Extrinsic and Intrinsic. Our work follows the following methodology: we start by studying the feasibility of adaptation of the F-measure for the evaluation of automatic summarization; after that, we define how to calculate the F-measure for a candidate summary. Text is presented with a term vector can be either a word or a phrase (with a binary-weighted or occurrence). To Determine to the exactitude of evaluation F-measure for automatic summarization by extraction calculates correlation with ROUGE Evaluation.