L'Evaluation de résumés informatifs

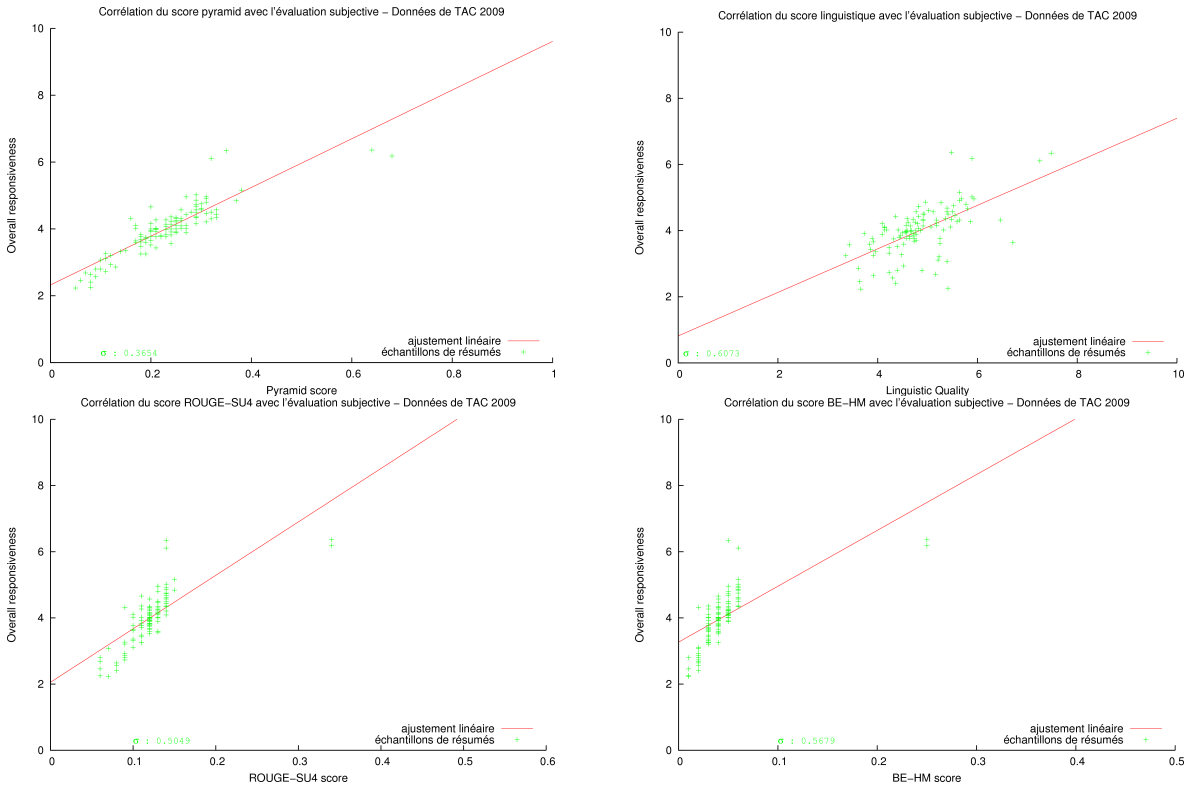
Sommaire

L'évaluation de résumés, qu'ils soient rédigés manuellement ou conçus par un système automatique, est une problématique à part entière. Un protocole d'évaluation de résumé doit faire intervenir deux aspects :

- le fond, ou les informations présentes dans le résumé ;
- la forme, ou la manière de présenter les informations.

Différentes mesures ont été proposées ces dernières années afin d'évaluer le fond, ou l'informativité des résumés de manière automatique. L'évaluation manuelle, telle que formulée dans les campagnes d'évaluation TAC [1] pose un problème évident de subjectivité. Cependant, cette évaluation manuelle s'avère être très corrélée aux résultats des évaluations manuelles guidées de l'informativité (méthode pyramide décrite plus bas), mais peu corrélée aux notes des résumés en qualité linguistique(*cf* fig.2.1). Cela signifie que les juges accordent plus d'importance au fond qu'à la forme des résumés automatiques, et que des mesures automatiques de l'informativité performantes sont essentielles à l'évaluation des résumés automatiques. Cependant, les évaluations de la qualité linguistique ne doivent pas être oubliées car la qualité linguistique est tout de même un critère important, qui permet de départager deux résumés sensiblement identiques du point de vue des informations qu'ils véhiculent.

1. Les évaluateurs devaient noter les résumés en fonction de leur réponse à la question : « Combien paieriez-vous pour ce résumé ? »



	qual. ling.	Pyramide	ROUGE-SU4	BE-HM
Coefficient de corrélation de Pearson	0.6210	0.8818	0.7586	0.6804

FIGURE 2.1.: Corrélation de l'évaluation subjective de résumés avec les évaluations linguistiques et informatives dans TAC 2009 : la mesure pyramide est la plus corrélée avec l'évaluation subjective. Viennent ensuite les mesures automatiques puis la mesure manuelle de la qualité linguistique.

2.1. ROUGE

Les mesures ROUGE, pour *Recall-Oriented Understudy for Gisty Evaluation*, ont été introduites par Lin dans (Lin, 2004). Ces mesures sont fondées sur la comparaison de n-grammes [2] entre un ou plusieurs résumés de référence et un résumé à évaluer. Il n'existe pas un unique résumé de référence, et il est donc essentiel de comparer les résumés automatiques à plusieurs résumés de référence établis manuellement afin d'obtenir des mesures plus précises de la qualité des résumés. Ces mesures nécessitent donc la rédaction de résumés de référence par un ou plusieurs experts au préalable de la mesure de qualité du résumé. Il en existe plusieurs variantes, que nous présentons ci-dessous.

2.1.1. ROUGE-n

Les métriques ROUGE-n sont fondées sur la comparaison simple de n-grammes. Une liste des n-grammes est établie pour chacun des résumés de référence et des résumés cible. Les n-grammes sont composés de n mots consécutifs. Par exemple, pour le texte « ROUGE est une métrique d'évaluation », la liste de n-grammes créée par ROUGE-2 sera « ROUGE est », « est une », « métrique d' », « d' évaluation ». Une fois la liste des n-grammes établie, le score ROUGE est calculé selon la formule en figure 2.1.1.

$$\frac{\sum_{r \in CRef} \sum_{n-grammes \in RC} card_{match}(n-grammes)}{\sum_{r \in CRef} \sum_{n-grammes \in RC} card(n-grammes)}$$

- $CRef$ est la collection des résumés de références,
- RC le résumé cible (le résumé à évaluer).

FIGURE 2.2.: Formule de calcul du score ROUGE-n : la formule revient à diviser le nombre de n-grammes communs entre le résumé à évaluer et les résumés de référence par le nombre total de n-grammes des résumés de référence.

Cette mesure présente un défaut majeur : l'ordre des mots n'influe pas toujours sur le résultat. Ainsi, l'exemple suivant, inspiré de la présentation par Lin de ROUGE au Workshop « Text Summarization Branches Out », montre à quel point les scores ROUGE-n peuvent être décorrélés de la réalité :

- Phrase 1 (référence) : Dr Jekyll tua Hide
- Phrase 2 : Dr Jekyll tue Hide
- Phrase 3 : Hide tue Dr Jekyll

2. Un n-gramme est une sous-séquence de n éléments construite à partir d'une séquence donnée.

Les scores ROUGE-n des phrases 2 et 3 seront similaires. En effet, les deux phrases ont les mêmes n-grammes en commun avec la phrase 1, à savoir (« Dr Jekyll », « Hide »).

2.1.2. ROUGE-L

Afin de pallier en partie un défaut de ROUGE-n, à savoir le fait que l'ordre des mots dans les phrases n'est pas pris en compte, (Lin, 2004) a introduit une autre mesure, ROUGE-L. Etant donné deux séquences S1 (référence) et S2, ROUGE-L est définie comme étant la plus longue sous-séquence commune (LCS) à S1 et S2 divisée par le nombre total d'éléments de S1. Ainsi, les scores des phrases 2 et 3 seront :

- phrase2 = (« Dr Jekyll Hide ») = 3/4
- phrase3 = (« Dr Jekyll ») = 2/4

Cependant, cette mesure n'est pas totalement satisfaisante. En effet, la phrase « Dr Jekyll a été tué par Hide. » obtiendrait avec cette mesure le même score que la « Dr Jekyll tue Hide. ».

2.1.3. ROUGE-SUn

ROUGE-SU, pour *skip-bigram and unigram ROUGE* prend en compte des bi-grammes à trou ainsi que les unigrammes des résumés de référence et du résumé cible. Les bi-grammes à trou (*skip-bigram*) sont définis dans (Lin, 2004) comme n'importe quelle paire de mots dans l'ordre de la phrase, séparés par une distance maximale arbitraire (le n). Ainsi, avec ROUGE-SU4, la phrase 1 comprend6 bi-grammes et 4 unigrammes : « Dr Jekyll », « Dr tua », « Dr Hide », « Jekyll tua », « Jekyll Hide », « tua Hide » et les 4 unigrammes qui composent la phrase. Les phrases 2 et 3 ont ainsi pour score respectivement : 6/10 et 4/10.

Cette mesure permet de rendre compte efficacement des relations de dépendance éloignées dans le texte. La campagne TAC 2008 l'a d'ailleurs confirmé puisque parmi les mesures ROUGE, c'est ROUGE-SU4 qui est la plus corrélée aux jugements humains.

2.2. BE-HM

Cette méthode d'évaluation automatique extrait des résumés de référence des *basic elements »* (BE). Les *Basic Elements* sont les unités sémantiques minimales (Hovy *et al.*,

		ROUGE-2		ROUGE-SU4	
	chaîne	n-grammes	score	n-grammes	score
phrase de référence	Résumer est un art difficile	Résumer est, est un, un art, art difficile		Résumer est, Résumer un, Résumer art, est un, est art, est difficile, un art, un difficile, art difficile + unigrammes	
phrase évaluée 1	Résumer n'est pas facile	Résumer n', n'est, est pas, pas facile	0	Résumer n', Résumer est Résumer pas, n' est, n' pas n' facile, est pas, est facile pas facile + unigrammes	3/15 = 0.2
phrase évaluée 2	Résumer est un art facile	Résumer est, est un, un art, art facile	0.75	Résumer est, Résumer un, Résumer art, est un, est art, est facile, un art, un facile, art facile + unigrammes	10/15 = 0.666

TABLE 2.1.: Exemple d'application des mesures ROUGE

2006) :

- soit la tête du constituant syntaxique majeur,
- soit la relation entre la tête et une unité qui en dépend, exprimée sous la forme d'un triplet (tête | modifieur | relation).

Par exemple, la phrase « Additional test flights are planned. » contient les BE suivants :

- flights|additional|mod
- flights|test|nn
- planned|flights|obj

L'inconvénient de cette méthode réside dans l'extraction automatique des BE des résumés de référence et des résumés à évaluer. D'une part, cette extraction introduit des erreurs puisque les systèmes sur lesquels elle est fondée ne sont pas totalement fiables. De plus, si l'on suppose qu'il existe une fonction permettant de passer de l'ensemble des BE des textes d'origine à l'ensemble des BE des résumés de référence, ou au moins d'approximer cet ensemble, alors cette évaluation favorise les systèmes intégrant le même type d'information. De plus, il a été montré dans (Dang et Owczarzak, 2008a) que cette mesure n'est pas aussi bien corrélée aux évaluations subjectives d'expert que les mesures classiques ROUGE présentées en §2.1.

2.3. Evaluation de résumés et théorie de l'information

La technique d'évaluation présentée dans (Lin *et al.*, 2006) est fondée sur l'hypothèse qu'étant donné un jeu de documents $D = \{d_1, d_2, ...d_i\}$, il existe une distribution de probabilités de paramètres fonction de θ_R qui génère un résumé de référence en fonction de D ; Résumer consiste alors à estimer ce θ_R. Les auteurs supposent que de la même manière, tout résumé réalisé automatiquement est généré par une distribution de probabilités de paramètres fonction de θ_A. Evaluer les résumés revient à calculer la divergence entre les deux distributions θ_R et θ_A. Les auteurs utilisent pour cela la mesure

Jensen-Shannon.

Cette technique d'évaluation a été utilisée avec succès sur le corpus DUC 2002, en résumé mono-document et multi-documents. Les classements des systèmes obtenus sont plus corrélés aux jugements humains que les mesures ROUGE-n pour le résumé multi-documents, et semblables aux mesures ROUGE-n en mono-document. Cependant, les données de DUC 2002 ne suffisent pas à l'évaluation d'une telle mesure : en effet, le nombre de systèmes qui ont participé (12 soumissions en multi-documents, 8 en mono-document). De plus, les mesures auxquelles les auteurs se comparent sont les mesures ROUGE-n, moins efficaces que les mesures ROUGE-SUn. Il est donc difficile de se faire une idée de la réelle efficacité de cette mesure d'évaluation.

2.4. La méthode Pyramide

Les systèmes automatiques d'évaluation de résumés sont tous confrontés au même problème : la reformulation. Les résumés de référence sont en effet écrits en respectant les règles de rédaction de résumé : identifier les informations principales et les reformuler. Par conséquent, le vocabulaire utilisé dans les résumés de référence ainsi que la structure de surface des phrases qui les composent varient des vocabulaire et structure des résumés automatiques, le plus souvent réalisés par des méthodes d'extraction.

L'impossibilité d'ajouter de la sémantique aux systèmes d'évaluation, sous peine de biaiser les résultats en favorisant les systèmes de résumés automatiques qui auront fait les mêmes choix, obligent à mettre au point des protocoles d'évaluation manuelle afin d'évaluer au mieux les résumés automatiques.

Nenkova *et al.* (2007) décrivent un tel protocole. A partir des résumés de référence, une liste de SCUs (*Single Content Units* ou unités de contenu) est établie. Elles sont classées et pondérées selon leur fréquence d'apparition dans les résumés de référence. La liste des SCUs prend alors la forme d'une pyramide (une structure hiérarchisée) reflétant l'importance des SCUs. L'étape suivante consiste à repérer dans les résumés à évaluer les SCUs présentes dans la pyramide. Un score est alors attribué aux résumés : la somme des poids des SCUs qu'ils contiennent, divisée par la somme des poids des SCUs des résumés de référence.

Ce type d'évaluation, peu automatisable, a le défaut d'être extrêmement coûteuse en temps d'expertise. En effet, apprendre à construire une pyramide de SCUs, de même que la construire, sont coûteux 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 2008 (Dang et Owczarzak, 2008a) et TAC 2009 (*cf fig. 2.1*).

En attendant des mesures plus précises d'évaluation entièrement automatique, il est donc indispensable de recourir à ce type d'évaluation manuelle lorsque les moyens humains s'y prêtent.

2.5. Évaluation de la forme

L'évaluation de la forme se rapproche de l'évaluation de la lisibilité d'un texte. Un texte lisible répond à différents critères. Il doit être à la fois :

- syntaxiquement correct ;
- sémantiquement cohérent ;
- articulé logiquement ;
- exempt de redondances.

En l'état actuel des recherches en TAL, cette partie de l'évaluation ne peut être réalisée dans sa totalité de manière automatique, surtout en ce qui concerne les évaluations de la cohérence sémantique et de l'articulation logique. Elle est par conséquent extrêmement coûteuse, puisqu'elle fait intervenir des jugements d'experts. Lors des campagnes d'évaluation DUC (Document Understanding Conference) et TAC (Texte Analysis Conference), les experts ont évalué la forme des résumés en se référant à une grille d'évaluation précise, présentée en figure 2.2.

Grammaticality	Grammaticalité	note de 1 à 10
Non-redundancy	Non-redondance	note de 1 à 10
Referential clarity	Auto-référentialité	note de 1 à 10
Focus	Mise en évidence des points principaux	note de 1 à 10 note de 1 à 10
Structure and Coherence	Structure et cohérence	note de 1 à 10

Table 2.2.: Grille d'évaluation linguistique des résumés des campagnes DUC et TAC

A partir de cette grille d'évaluation, les experts attribuent une note de qualité linguistique à chacun des résumés automatiques. Il est intéressant de noter qu'au cours de la campagne d'évaluation TAC 2009, les résumés rédigés par les experts ont obtenu une note linguistique de 8.8/10, tous experts confondus. Malheureusement, les résultats publiés mentionnent uniquement la note linguistique globale ; les notes détaillées des différents critères de la grille d'évaluation n'y apparaissent pas, ce qui nous empêche d'aller plus loin dans la compréhension de ces notes. Cependant, cette grille d'évaluation paraît adéquate. En effet, elle rend parfaitement compte des critères de lisibilité énoncés plus haut.

2.6. Conclusion

Nous avons effectué un panorama des techniques d'évaluation de résumés automatiques. Aucune solution actuelle n'est complètement satisfaisante. En effet, les méthodes entièrement automatiques (ROUGE, BE-HM) manquent de précision. De plus, il est possible d'adapter des systèmes de résumé automatique de manière à ce qu'ils obtiennent de meilleurs scores avec une technique automatique sans pour autant augmenter la qualité du résumé. Cependant l'utilisation de méthodes manuelles qui ne font pas seulement intervenir un expert pour la rédaction d'un ou plusieurs résumés de référence, mais également pour la construction d'un modèle d'évaluation depuis ces résumés, et la confrontation des résumés automatiques à ce modèle, sont particulièrement coûteuses en temps. Une évaluation précise ne peut donc, en l'état actuel des avancées dans ce domaine, être réalisée sur des données trop importantes. Des efforts doivent donc être menés par la communauté afin de développer des méthodes d'évaluation précises et peu coûteuses, qui pourront supporter le passage à l'échelle.