Une méthode d'évaluation des résumés basée sur la combinaison de métriques automatiques et de complexité textuelle

Samira Walha Ellouze, Maher Jaoua, Lamia Hadrich Belguith

ANLP Research Group, Laboratoire MIRACL, Route de l'aéroport Km 4, 3018, Sfax, Tunisie

ellouze.samira@gmail.com, Maher.Jaoua@fsegs.rnu.tn,
l.belguith@fsegs.rnu.tn

RÉSUMÉ

Cet article présente une méthode automatique d'évaluation du contenu des résumés automatiques. La méthode proposée est basée sur une combinaison de caractéristiques englobant des scores de contenu et d'autres de complexité textuelle et ce en s'appuyant sur une technique d'apprentissage, à savoir la régression linéaire. L'objectif de cette combinaison consiste à prédire le score manuel PYRAMID à partir des caractéristiques utilisées. Afin d'évaluer la méthode présentée, nous nous sommes intéressés à deux niveaux de granularité d'évaluation : la première est qualifiée de Micro-évaluation et propose l'évaluation de chaque résumé, alors que la deuxième est une Macro-évaluation et s'applique au niveau de chaque système.

ABSTRACT

An evaluation summary method based on combination of automatic and textual complexity metrics

This article presents an automatic method for evaluating content summaries. The proposed method is based on a combination of features encompassing scores of content and others of textual complexity. This method relies on a learning technique namely the linear regression. The objective of this combination is to predict the PYRAMID score from used features. In order to evaluate the presented method, we are interested in two levels of granularity evaluation: the first is named Micro-evaluation and proposes an evaluation of each summary, while the second is called Macro-evaluation and it applies at the level of each system.

MOTS-CLÉS : Evaluation intrinsèque, évaluation du contenu, résumé automatique, complexité textuelle, régression linéaire.

KEYWORDS: Intrinsic evaluation, content evaluation, automatic summary, textual complexity, linear regression.

1 Introduction

L'évaluation d'un résumé est une tâche importante et nécessaire. Elle permet de quantifier la qualité informationnelle et linguistique d'un résumé et peut être de deux types : extrinsèque ou intrinsèque. L'évaluation extrinsèque permet d'évaluer la qualité du résumé par rapport à une tâche annexe telle que la classification des documents et l'indexation, alors que l'évaluation intrinsèque permet d'évaluer la qualité globale du résumé d'une manière manuelle ou automatique. Il est à noter que l'évaluation manuelle est une tâche difficile et coûteuse vu qu'elle nécessite un temps important et une expertise du domaine du thème du

texte source. Pour cette raison, plusieurs mesures d'évaluation automatique ont été développées, telles que ROUGE, BE, AutoSummENG, etc. Afin d'évaluer l'exactitude des mesures automatiques, on effectue généralement une comparaison entre ces mesures et les scores d'évaluation obtenus manuellement. Pour effectuer cette comparaison, la compagnie d'évaluation TAC[1] a proposé diverses mesures de corrélations (i.e. Pearson, Spearman). La plupart des mesures évaluées par la conférence TAC se basent sur l'évaluation de la pertinence du contenu. Toutefois, un résumé avec un contenu pertinent peut être illisible. Afin d'encourager les chercheurs à évaluer la lisibilité d'un résumé, la session TAC 2011(Owczarzak et Dang, 2011) a ajouté un nouvel objectif à la tâche d'évaluation automatique des résumés qui consiste à évaluer la lisibilité des résumés.

Dans ce contexte, nous proposons dans ce travail une méthode d'évaluation basée sur la combinaison de plusieurs mesures d'évaluation, à savoir des mesures de contenu et de lisibilité textuelle. Cet article s'articule autour de trois parties. La première partie présente un panorama des principales méthodes d'évaluation intrinsèques utilisées dans le domaine du résumé automatique. La deuxième partie décrit la méthode proposée qui opère par combinaison linéaire des caractéristiques statistiques et linguistiques des résumés. Quant à la dernière partie, elle présente les résultats de nos expérimentations.

2 Panorama des mesures intrinsèques

Les premières évaluations dans le domaine du résumé automatique sont effectuées par des juges humains. Ces juges évaluent un résumé en répondant à des questions sur la cohérence, la couverture, la pertinence, etc. Cette façon d'évaluer nécessite des ressources humaines importantes. De même, l'évaluation humaine est subjective puisqu'elle varie d'un évaluateur à un autre. D'ailleurs, elle peut varier pour un même évaluateur lors de deux évaluations séparées dans le temps. Malgré tous ces inconvénients l'évaluation par des juges humains est utilisée par plusieurs mesures d'évaluation telles que « Overall Responsiveness » qui mesure une combinaison de contenu et de qualité linguistique. Dans ce qui suit, nous donnons un aperçu sur les méthodes les plus utilisées pour l'évaluation manuelle et automatique.

2.1 Méthodes manuelles

Afin de remédier aux inconvénients de l'évaluation des jugements humains, la compagnie d'évaluation DUC a utilisé l'interface SEE (Lin, 2001) qui permet aux juges humains d'évaluer manuellement le contenu et la qualité linguistiques (i.e., la grammaticalité, la cohésion, la cohérence, etc.) d'un résumé. A partir de l'année 2005, la compagnie DUC a commencé à utiliser la méthode manuelle PYRAMID (Nenkova et Passonneau, 2004). Cette méthode se base sur l'identification des unités minimales sémantiques appelées SCUs (Summarization Content Units). Afin de construire la pyramide, les annotateurs identifient manuellement les SCUs des résumés de référence. La position d'un SCU dans la pyramide diffère selon son nombre d'occurrences dans les résumés de référence. Il s'agit, ensuite, d'évaluer les résumés candidats en dégageant les SCUs de chaque résumé candidat, puis en les comparants avec les SCUs de la pyramide. La méthode PYRAMID a pu limiter le désaccord entre les annotateurs en leur donnant une flexibilité en matière de définition des SCUs. Mais le guide d'annotation

[1]Text analysis Conference http://www.nist.gov/tac

lui-même peut être soumis à des critères d'évaluation différents selon la tâche visée.

2.2 Méthodes automatiques

A cause des difficultés rencontrées lors de l'évaluation manuelle, plusieurs recherches se sont orientées vers l'évaluation automatique. ROUGE, proposée par (Lin, 2004), est l'une des premières mesures automatiques qui ont été conçues pour l'évaluation des résumés. Elle se fonde sur une méthode à base du chevauchement des N-grammes du résumé candidat avec un ou plusieurs résumés de référence. (Hovy et al., 2006) ont introduit la mesure BE (Basic Elements) permettant de faire la correspondance entre des unités syntaxiques courtes appelées BE. Dans un travail plus récent, (Giannakopoulos et al., 2008) ont introduit la mesure AutoSummENG permettant de représenter un résumé sous forme de graphe de n-grammes (de caractères ou de mots) et de faire la comparaison entre deux graphes. D'autres mesures d'évaluation n'utilisant pas des résumés de référence ont aussi été proposées par (Louis and Nenkova, 2009) et (Torres-Moreno et al., 2010). Ces mesures permettent de comparer chaque résumé candidat aux documents sources en utilisant la mesure de divergence de Jensen-Shannon.

Des nouvelles mesures telles que ROSE (Conroy et Dang, 2008) et Nouveau-ROUGE (Conroy et al., 2011) ont mis en jeu la combinaison de plusieurs variantes de ROUGE afin de prédire le score de PYRAMID ou de Overall Responsiveness. D'autres travaux se sont intéressés aux métriques d'évaluation de la qualité linguistique. Dans ce cadre, (Pilter et al., 2010) ont évalué les cinq propriétés linguistiques utilisées dans TAC en combinant différents types de caractéristiques telles que la grille d'entité de (Barzilay et Lapata, 2008), le cosinus similarité entre les représentations vectorielles des phrases adjacentes, etc. Les travaux les plus récents, tel celui de (Conroy et al, 2010), ont évalué le contenu et la qualité grammaticale en utilisant une combinaison de caractéristiques. Concernant les caractéristiques de contenu, (Conroy et al, 2010) utilisent les scores de ROUGE pour les résumés de base et les scores de Nouveau-ROUGE pour les résumés de mise à jour. Dans un travail ultérieur, (Conroy et al., 2011) et (Rankel et al., 2012) ont combiné des caractéristiques de contenu (à base de bi-grammes) et d'autres linguistiques. A l'opposé des travaux de Conroy, (Lin et al., 2012) ont combiné une mesure d'évaluation de la traduction automatique à base de N-grammes ainsi qu'une mesure de cohérence à base de grille d'entité afin de prédire Overall Responsiveness.

3 Méthode proposée

En se basant sur les travaux réalisés, nous avons constaté que les méthodes utilisant des techniques d'apprentissage sont les plus adaptées pour obtenir des résultats proches de la méthode manuelle PYRAMID. C'est pour cela que nous avons proposé une méthode à base d'apprentissage, permettant de prédire la mesure PYRAMID. Donc, nous avons développé un modèle de régression linéaire qui combine des mesures de contenu ROUGE, BE et AutoSummENG, des mesures linguistiques telles que la densité des mots fonctionnels ainsi que des mesures de complexité textuelle à base du nombre des phrases, nombre de mots par phrase, etc. Ainsi, l'équation d'estimation du score PYRAMID s'écrit :

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

Le problème de régression linéaire est donc exprimé comme un ensemble de caractéristiques et le score PYRAMID qui leur correspond. Par la suite, nous déterminons un

vecteur w qui maximise la corrélation tel que :$w = argmax\ \rho(\sum_{j=1}^{n} x_{ij} w_j\ , y_i)$

Avec x_{ij}la valeur de la j[ième] caractéristique pour le système i (respectivement pour un résumé i) lors de la macro-évaluation (respectivement lors de la micro-évaluation) ; y_i le score PYRAMID du système i (respectivement du résumé i) lors de la macro-évaluation (respectivement lors de la micro-évaluation) avec i allant de 1 à m et j allant de 1 à n et ρ la corrélation de Pearson. Nous utilisons la méthode des moindres carrés pour minimiser la somme des écarts au carré entre le score PYRAMID et le score estimé de PYRAMID. Donc l'équation de minimisation s'écrit : $min \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$

4 Caractéristiques

Les caractéristiques utilisées par notre méthode sont choisies de sorte que leur combinaison corrèle le maximum avec le score PYRAMID.

4.1 Caractéristiques du contenu

A partir des résultats de corrélation obtenus dans TAC 2008 (Dang et Owczarzak, 2008), nous remarquons que les mesures standards ROUGE-2, ROUGE-SU4 et BE-HM[2], ainsi que la mesure AutoSummENG disposent d'une corrélation élevée avec la mesure PYRAMID. Pour cela, nous avons utilisé principalement ces quatre mesures comme caractéristiques d'évaluation du résumé. De plus, nous avons ajouté d'une part les mesures ROUGE-3 et ROUGE-4 vu qu'elles prennent en considération des contextes larges, permettant ainsi de capturer des caractéristiques linguistiques tels certains phénomènes grammaticaux, et d'autre part ROUGE-1 vu qu'elle présente un bon indicateur de la pertinence du contenu d'un résumé.

4.2 Caractéristiques linguistiques

PYRAMID est une méthode manuelle basée sur l'extraction des SCUs. Un juge humain ne peut pas identifier les SCUs dans un résumé n'ayant pas une bonne qualité linguistique. Donc, un résumé avec une mauvaise qualité linguistique ne peut pas avoir un bon score PYRAMID. Ainsi, il est intéressant d'inclure des mesures linguistiques pour garantir une meilleure prédiction du score PYRAMID. Nous citerons par la suite des caractéristiques linguistiques permettant d'influencer la qualité du résumé.

4.2.1 Densité des mots fonctionnels

La densité de diverses catégories de mots fonctionnels peut nous informer sur la cohésion d'un texte. En fait, selon (Halliday et Hasan, 1976), le concept de cohésion englobe les phénomènes (i.e. les connecteurs de discours, les dispositifs de coréférence, etc.) permettant de relier les phrases entre elles. Par exemple, les connecteurs du discours tels que « but », « and », « while » permettent de relier des événements exprimés par différentes phrases. Vu que plusieurs mots fonctionnels représentent des dispositifs de coréférence ou des connecteurs de discours, nous avons décidé de calculer la densité des catégories suivantes des mots fonctionnels : les déterminants (DET), les conjonctions de coordination (CC), les

[2] BE-HM utilise la tête et le modificateur seulement.

prépositions, les conjonctions de subordination (PCS) et les pronoms personnels (PRP). La densité de chacune des catégories précédentes représente le ratio entre le nombre de mots présentant l'une des catégories et le nombre total des mots dans le résumé. Nous détectons les mots fonctionnels à l'aide de l'étiqueteur morphologique "Stanford Postagger[3]".

4.2.2 Mesures de complexité textuelle

L'analyse de lisibilité nous permet de déterminer si un texte est facile à comprendre ou non ; autrement dit, elle permet d'indiquer la complexité du texte. Les mesures de lisibilité utilisées dans ce travail sont basées sur le nombre de phrases, de mots, de caractères, de syllabes et/ou de mots complexes dans un résumé. Ces mesures sont :

- La mesure Gunning Fog Index (GFI) qui indique la lisibilité d'un texte rédigé en anglais. Plus précisément, c'est un indice permettant d'indiquer les années de scolarité nécessaires pour comprendre le texte lors d'une première lecture. La formule utilisée est la suivante : $score = 0{,}4(\mathrm{LMP} + \mathrm{PMC})$, où LMP représente la longueur moyenne d'une phrase et PMC représente le pourcentage des mots avec trois syllabes ou plus.
- La mesure Flesch Reading Ease (FRE) : elle permet de prédire la difficulté des documents à lire pour l'adulte. Sa formule s'écrit : $score = 206{,}835 - (1{,}015 * \mathrm{LMP}) - (84{,}6 * \mathrm{MSM})$, où MSM représente le nombre moyen de syllabes par mot.
- La mesure Flesch-Kincaid Index (FKI) : elle permet de juger le niveau de lisibilité des textes et des livres anglais, c'est-à-dire qu'elle indique la difficulté de compréhension lors de la lecture de ces textes et de ces livres. Cette mesure est intégrée dans l'outil Word de Microsoft. Sa formule est la suivante : $score = 0{,}39 * \mathrm{LMP} + 11{,}8 * \mathrm{MSM} - 15{,}59$
- Nombre de phrases (NbPh) : nous utilisons l'indice employé par [Rankel et al., 2012] qui est égal à -log(nombre de phrases).

4.2.3 Pénalité de dépassement de longueur

En examinant les résultats des différents systèmes qui ont participé à la conférence TAC, nous remarquons que les résumés ayant subi une troncature à la fin (à cause du dépassement de la taille maximale autorisée par la conférence) ont été pénalisés dans leur score de réactivité globale et dans leur score de qualité linguistique. Pour cela, nous avons ajouté comme caractéristique une mesure de pénalité de dépassement de longueur(PDL). Cette mesure est égale au rapport entre le nombre de mots dans un résumé et la taille maximale des résumés TAC (maximum 100 mots).

5 Evaluation

L'évaluation de la nouvelle métrique est basée sur l'étude de sa corrélation avec la métrique PYRAMID. Nous utilisons 3 mesures de corrélation, à savoir la corrélation de Pearson, celle de Spearman et celle de Kendall. Nous utilisons le corpus de TAC 2008 pour évaluer notre métrique. Ce corpus comporte 48 thèmes et 58 systèmes de résumés. Pour chaque thème, il existe 20 documents triés en ordre chronologique. Chaque système produit un résumé de base construit à l'aide des 10 premiers documents seulement et un autre résumé de mise à jour construit à partir des 10 documents suivants. Un résumé de mise à jour décrit les

[3]Cet étiqueteur fournit des inférences bidirectionnelles. (http://www-nlp.stanford.edu/software/tagger.shtml)

événements évolutifs, c'est à dire les nouveaux événements apportés par les 10 derniers documents par rapport aux événements décrits dans les 10 premiers documents. Au total, chaque système a produit 96 résumés (48 résumés de base et 48 résumés de mise à jour). Nous examinons le pouvoir prédictif de nos caractéristiques sur deux niveaux : niveau résumé (Micro-évaluation) et niveau système (Macro-évaluation). Dans les deux niveaux, nous employons la méthode de validation croisée « k-fold cross validation » avec k égal à 10.

5.1 Micro-évaluation

Dans cette partie, nous étudions la capacité de prédiction des caractéristiques utilisées au niveau d'une micro-évaluation. Autrement dit, nous effectuons une évaluation niveau résumé dans laquelle nous prenons le score de chaque résumé dans une entrée à part. Nous avons réalisé une expérimentation pour chaque tâche d'évaluation des résumés. Les caractéristiques utilisées dans la Micro-évaluation sont présentées dans la table suivante :

Évaluation du résumé	Caractéristiques
de base	R1, R2, R3, R-SU4, BE, AutoSummENG, NbPh, densité(DET), FKI, GFI
de mise à jour	R1, R2, R3, R4, R-SU4, BE, AutoSummENG, NbPh, PDL, Densité(DET), Densité(PCS), GFI

TABLE 1–Caractéristiques utilisées dans la tâche des résumés de base et de mise à jour au niveau de la Micro-évaluation

A partir de la table 2 et dans les deux niveaux d'évaluation, nous constatons que la corrélation de notre expérimentation avec PYRAMID n'est pas assez élevée, bien qu'elle soit plus grande que celle des standards (ROUGE-SU4, ROUGE-2, BE) avec PYRAMID.

	Résumé de base			Résumé de mise à jour		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
ROUGE-2	0,4646	0,4855	0,3361	0,5645	0,6033	0,4252
ROUGE-SU4	0,4942	0,5070	0,3531	0,6013	0,6359	0,4505
BE	0,3796	0,4122	0,2831	0,5391	0,5968	0,4213
Notre expérimentation	**0,5960**	**0,5901**	**0,4185**	**0,6528**	**0,6760**	**0,4873**

TABLE 2–Corrélation avec PYRAMID dans TAC 2008 tâche d'évaluation des résumés de base et des résumés de mise à jour, micro-évaluation (p-value<2,2e-16)

5.2 Macro-évaluation

Dans cette partie, nous effectuons une Macro-évaluation, c'est-à-dire une évaluation au niveau système. Dans ce type d'évaluation, nous calculons la moyenne des scores obtenus par chaque système. Pour chaque tâche d'évaluation, nous avons effectué une expérimentation. La table 3 donne un aperçu sur les caractéristiques utilisées dans chaque tâche.

Évaluation résumé	Caractéristiques
de base	R1, R2, NbPh, PDL, GFI, densité(DET), densité(CC), FKI, GFI, FRE
de mise à jour	R1, R3, R4, BE, AutoSummENG, PDL, Densité(CC), Densité(PRP), GFI, FKI

TABLE 3 –Caractéristiques utilisées dans la tâche des résumés de base et de mise à jour au niveau de la Macro-évaluation

La table 4 donne les coefficients de corrélation du score PYRAMID avec les mesures standards : ROUGE-2, ROUGE-SU4 et BE, les expérimentations décrites dans la table 3 ainsi que les mesures Nouveau-Rouge-2, Nouveau-Rouge-SU4 qui sont réalisées par (Conroy et al., 2011) pour évaluer les résumés de mise à jour au niveau de la macro-évaluation seulement.

	Résumé de base			Résumé de mise à jour		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
ROUGE-2	0,8981	0,9095	0,7611	0,9366	0,9415	0,8000
ROUGE-SU4	0,8780	0,8859	0,7340	0,9174	0,9310	0,7842
BE	0,9045	0,9022	0,7319	0,9398	0,9376	0,7951
Nouveau-ROUGE-R2	-	-	-	0,9525	0,9434	0,8085
Nouveau-ROUGE-SU4	-	-	-	0,9359	0,9339	0,7908
Notre expérimentation	**0,9700**	**0,9552**	**0,8253**	**0,9704**	**0,9684**	**0,8582**

TABLE 4–Corrélation avec PYRAMID dans TAC 2008 tâche d'évaluation des résumés de base et de mise à jour, macro-évaluation (p-value< 4,4 e-16)

En examinant la table 4, nous apercevons que nos expérimentations donnent une bonne corrélation avec PYRAMID. Nous constatons également que notre expérimentation est meilleure que les mesures standard utilisées par TAC ainsi que les deux variantes de la mesure Nouveau-ROUGE qui ont été destinées à l'évaluation des résumés de mise à jour.

6 Conclusion

Dans cet article, nous avons présenté une méthode d'évaluation du contenu d'un résumé, en utilisant une combinaison de caractéristiques linguistiques et de contenu. La combinaison de ces caractéristiques est réalisée à l'aide d'une régression linéaire. Nous avons utilisé comme caractéristiques : les mesures automatiques les plus corrélées avec PYRAMID dans TAC 2008, les mesures de lisibilité les plus utilisées par les sites internet et les outils de traitement de texte ainsi que la densité de quelques catégories de mots fonctionnels.

En examinant les résultats obtenus, nous constatons que nous pourrions les améliorer davantage à travers l'intégration de nouvelles mesures à base de ROUGE en utilisant l'ontologie WordNet. Cette ontologie nous permettrait de résoudre le problème de l'emploi des mots synonymes dans les résumés. De même, nous pourrions utiliser d'autres caractéristiques linguistiques telles que la grille d'entités, utilisée par (Barzilay et Lapata, 2008) pour mesurer la cohérence du résumé.

Références

BARZILAY, R. and LAPATA, M. (2008). Modeling Local Coherence: An Entity-based Approach. *In Computational Linguistics Journal*, Vol: 34 No: 1, pages 1-34.

CONROY, J. M., SCHLESINGER, J. D. and O'LEARY, D. P. (2011). Nouveau-ROUGE: A Novelty Metric for Update Summarization. *In Computational Linguistics journal*, Vol: 37 No: 1, pages 1-8.

CONROY, J. M., SCHLESINGER, J. D., RANKEL, P. A., and O'LEARY, D. P. (2010). Guiding CLASSY toward More Responsive Summaries. *In proceedings of the Text Analysis Conference*.

CONROY, J. M. and TRANG DANG, H. (2008). Mind the Gap: Dangers of Divorcing Evaluations of

Summary Content from Linguistic Quality. *In proceedings of COLING 2008*, pages 145-152.

DANG, H. T. and OWCZARZAK, K. (2009). Overview of TAC 2009 summarization track. *In proceedings of the Text Analysis Conference*.

DANG, H. T. and OWCZARZAK, K. (2008). Overview of the TAC 2008 Update Summarization Task. *In proceedings of the Text Analysis Conference*.

GIANNAKOPOULOS, G. and KARKALETSIS, V. (2010). Summarization system evaluation variations based on n-gram graphs. *In the proceedings of TAC 2010 Workshop*.

GIANNAKOPOULOS, G., KARKALETSIS, V., VOUROS, G. A. and STAMATOPOULOS, P. (2008). Summarization system evaluation revisited: N-gram graphs. *TSLP journal,* Vol: 5 No: 3.

HALLIDAY, M. A. K. and HASAN, R. (1976). *Cohesion in English*. Longman (Londres).

HARNLY, A., NENKOVA, A., PASSONNEAU, R. and RAMBOW, O. (2005). Automation of Summary Evaluation by the Pyramid Method. *In proceedings of RANLP*, pages 226-233.

HOVY, E., LIN, C., ZHOU, L. and FUKUMOTO, J. (2006). Automated Summarization Evaluation with Basic Elements. *In proceedings of the 5th Conference on Language Resources and Evaluation*.

LIN, C. (2001). Summary Evaluation Environment. http://www.isi.edu/~cyl/SEE.

LIN, C. (2004). ROUGE: a package for automatic evaluation of summaries. *In proceedings of the ACL 2004 Workshop on Text Summarization Branches Out (WAS 2004)*, pages 74–81.

LIN, Z., LIU, C., NG, H. T. and KAN, M. (2012). Combining Coherence Models and Machine Translation Evaluation Metrics for Summarization Evaluation. *In proceedings of ACL (1)*, pages 1006-1014.

LIN, Z., NG, H. T. and KAN, M. (2011). Automatically Evaluating Text Coherence Using Discourse Relations. *In proceedings of ACL 2011*, pages 997-1006.

LOUIS, A. and NENKOVA, A. (2009). Automatically Evaluating Content Selection in Summarization without Human Models. *In proceedings of EMNLP 2009*, pages 306-314.

NENKOVA, A. and PASSONNEAU, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. *In proceedings of HLT-NAACL 2004*, pages 145-152.

OWCZARZAK, K. and DANG, H. T. (2011). Overview of the TAC 2011 summarization track: Guided task and AESOP task. *In proceedings of the Text Analysis Conference*.

PITLER, E., LOUIS, A. and NENKOVA, A. (2010). Automatic Evaluation of Linguistic Quality in Multi-Document Summarization. *In proceedings of ACL 2010*, pages 544-554.

RANKEL, P. A., CONROY, J. M. and SCHLESINGER, J. D. (2012). Better Metrics to Automatically Predict the Quality of a Text Summary. *Algorithms journal*, No: 4, pages 398-420.

TORRES-MORENO, J. M., SAGGION, H., DA CUNHA, I., San-Juan, E. and VELAZQUEZ-MORALES, P. (2010). Summary Evaluation With and Without References. *Polibits ISSN1870-9044*, pages 13-19.