Un système de résumé de textes en arabe

F. S. Douzidia, G. Lapalme

RALI – Département d'informatique et de recherche opérationnelle,
Université de Montréal
C.P. 6128, Succ Centre-Ville Montréal, Québec, Canada H3C 3J7
{douzidif, lapalme}@iro.umontreal.ca
Télécopieur : (001) 514-343-5834

Résumé

Cet article propose une méthode de production de résumés pour les textes arabes. Nous avons adapté à la langue arabe les techniques de résumé automatique qui ont prouvé leur efficacité pour d'autres langues et qui donnent des résultats satisfaisants au niveau informationnel. Grâce aux techniques de compression que nous avons introduites, nous avons pu montrer, lors d'une compétition d'évaluation de résumés automatiques, que les traductions des résumés produits par notre système étaient meilleures que des résumés produits à partir de textes traduits.

Mots Clés

Résumé automatique, Traitement automatique de la langue arabe, Extraction de l'information, traduction arabe.

Abstract

This paper presents a method for producing summaries for Arabic texts. Our approach adapts to Arabic texts the techniques of automatic summarization, which have proven their effectiveness for other languages and given relevant results in terms of informational level. Combined with compression techniques, we have shown in a competition of evaluation of automatic summaries, that the translations of summaries we produced were better than summaries of translated texts.

Keywords

Automatic summarization, Arabic language processing, information extraction, Arabic translation.

1 Introduction

Le but d'un résumé automatique de texte est de produire une représentation abrégée d'un document ou, dans certains cas, de plusieurs documents. Le résumé de texte automatique peut être classifié en deux approches : abstraction et extraction. Alors que l'abstraction exige l'utilisation du traitement de langage naturel, incluant des grammaires et des lexiques pour l'analyse syntaxique et la génération, l'extraction choisit des parties appropriées (des phrases, des paragraphes, etc.) du texte original et les enchaîne sous une forme plus courte [1].

La plupart des travaux récents dans le domaine du résumé sont basés sur l'extraction, même si la lecture des résumés par extraction puisse être difficile en raison du manque de cohérence.

L'évaluation DUC 2004 (Document Understanding Conference) a décidé d'inclure une tâche de résumé automatique en anglais de documents arabes. L'approche proposée lors de la compétition est d'extraire des résumés à partir de textes traduits automatiquement de l'arabe à l'anglais. Les documents arabes sont issus de l'Agence France Presse (A.F.P.) Arabic Newswire (1998, 2000-2001). Ils sont traduits vers l'anglais par deux systèmes automatiques: un de l'Institut des Sciences de l'Information (ISI) de Université de la Californie du Sud et un développé par IBM. Une étude préliminaire sur la qualité d'un échantillon de documents traduits par ces systèmes a révélé les problèmes suivants :

- Les textes en anglais sont difficilement compréhensibles sans avoir recours à l'original en arabe.
- Les systèmes de traduction ont souvent omis des informations importantes. Par exemple la traduction suivante de la troisième phrase du document afa20030101.5900 produite par ISI, *The Agency said that Ibrahim, in the event at the level of cooperation and trade between Iraq and Saudi Arabi*[1]. Le verbe *appreciated* a été omis après le mot *event*, même là le paragraphe reste difficile à comprendre.
- Ces systèmes traduisent souvent le même mot de deux manières différentes, ce qui pourrait influencer les résultats sur les résumés. Par exemple, IBM a traduit le mot *maisons*-منازل dans le document afa20030102.4200 par *home* dans la première phrase et par *workers* dans la deuxième phrase.

Nous avons donc plutôt proposé aux organisateurs de DUC 2004, qui ont accepté, de prendre une autre voie: travailler en arabe pour effectuer le résumé pour ne finalement traduire en anglais que le résumé. Nous avons donc moins de texte à traduire automatiquement.

Cette approche nous a permis d'obtenir de meilleurs résultats, car on travaille sur les documents qui n'ont pas subi d'altération de traduction, surtout si on considère les ambiguïtés que génèrent les traducteurs automatiques.

Dans cet article nous allons d'abord décrire brièvement les principaux modules du système Lakhas[2] que nous avons développé. Nous présentons brièvement ensuite les extensions faites à Lakhas et les résultats que nous avons obtenus lors de cette évaluation.

---

[1] Texte original: وقالت الوكالة ان ابراهيم رحب في المناسبة بمستوى التعاون والتبادل التجاري بين العراق والسعودية

[2] Transcription stricte de *résumer* en arabe

2 Architecture de LAKHAS

La mise en œuvre fonctionnelle de LAKHAS est représentée à la Figure 1. Elle repose sur la segmentation à différents niveaux ainsi que sur le calcul des poids afin de permettre la génération de résumé. Nous décrivons maintenant brièvement les modules selon la numérotation de la Figure 1.

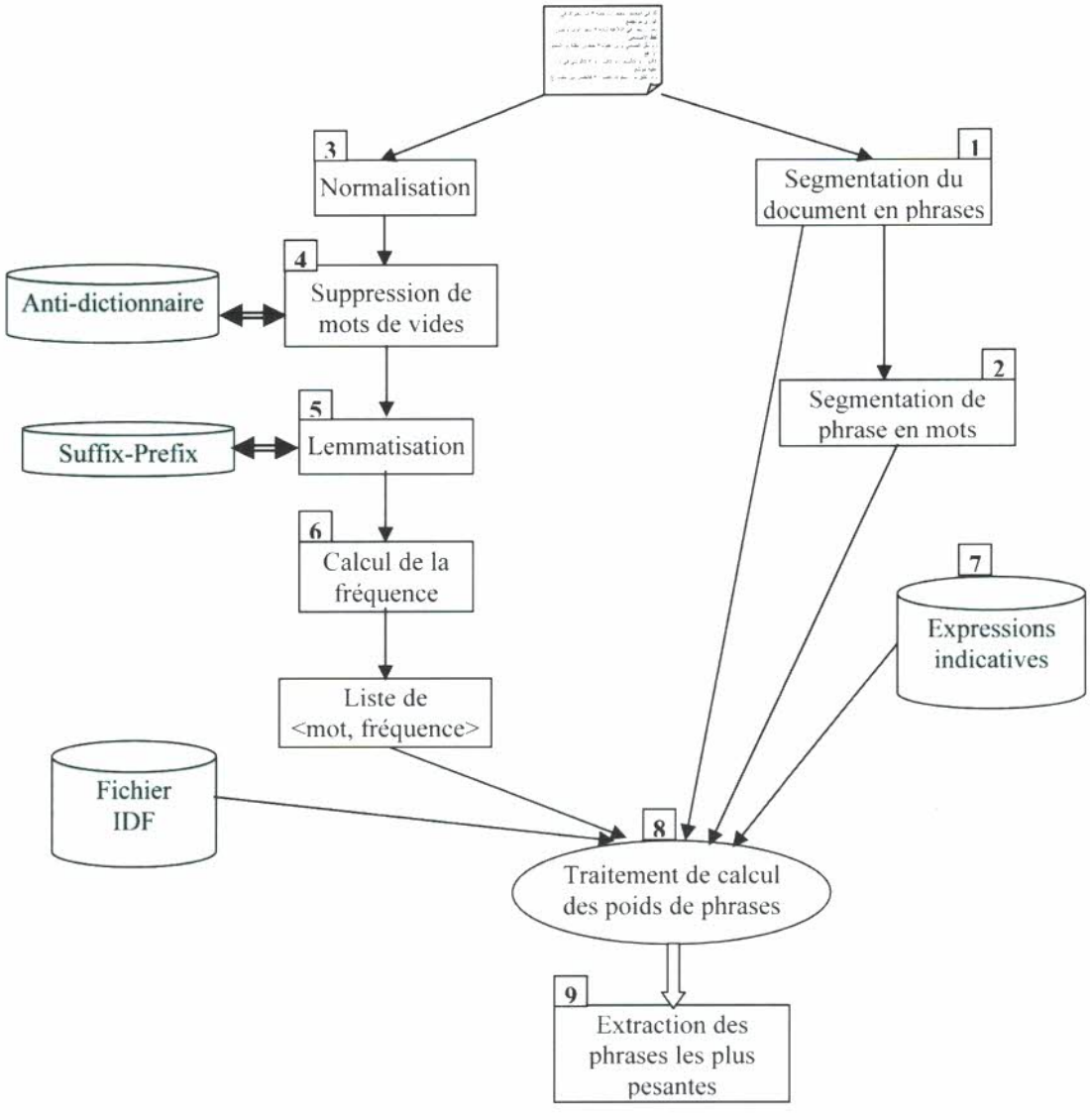

Figure 1 : Schéma global de Lakhas, les numéros des modules font références aux items de la section 2.

1. ***Segmentation du document en phrases*** identifie chaque phrase du document ainsi que sa position. Dans le corpus chaque phrase du document est délimité par <P> et </P>, et le titre est délimité par les balises <HEADLINE> et </HEADLINE>. Il reçoit en entrée le document et retourne le résultat sous forme de liste de phrases. Dans cette liste, le titre du document est inséré en tête de liste dans le but de l'utiliser pour le calcul du poids.

2. ***Segmentation des phrases en mots,*** appelé aussi tokenisation sépare chaque phrase en une séquence de mots, en détectant les délimiteurs de mot tels que l'espace ou la

ponctuation. Ceci permet de retourner une liste de mots avec leur fréquence par rapport à la phrase.

3. ***Normalisation*** transforme le document dans un format standard plus facilement manipulable. Avant la lemmatisation, le document est normalisé comme suit [2]:

   - Suppression des caractères spéciaux et les chiffres
   - Remplacement de ! , أ et إ avec ا
   - Remplacement de la lettre finale ي avec ى
   - Remplacement de la lettre finale ة avec ه

   Cette étape est nécessaire à cause des variations qui peuvent exister lors de l'écriture d'un même mot.

4. ***Suppression des mots vides*** consiste à éliminer tous les mots non significatifs. On compare chaque mot reconnu avec un des éléments dans l'antidictionnaire qui contient tous les mots non-significatifs. Si un mot en fait partie, il ne sera pas pris en considération pour le calcul de sa fréquence.

5. ***Lemmatisation***
   C'est une tâche délicate du fait que l'arabe est une langue flexionnelle et fortement dérivable [3]; l'absence des diacritiques crée une ambiguïté et donc exige des règles morphologiques complexes, de plus la capitalisation n'est pas employée dans l'arabe ce qui rend difficile l'identification des noms propres, des acronymes et des abréviations. Pour résoudre l'ambiguïté, Aljlayl et Frieder montrent que la lemmatisation légère (approche basée sur suppression de suffixe et de préfixe) surpasse significativement celle basée sur détection de racine dans le domaine de recherche d'information [4]. Pour notre cas nous avons considéré la lemmatisation légère qui consiste à déceler si des préfixes ou suffixes ont été ajoutés au mot [5]. Puisque la plupart des mots arabes ont une racine à trois ou quatre lettres, le fait de garder le mot au minimum à trois lettres va permettre de préserver l'intégrité du sens du mot.

   Nous utilisons la liste de préfixes et de suffixes proposée par [6] voir Tableau 1. Plusieurs d'entre eux ont été utilisés par [7] pour la lemmatisation de mots arabes; ils ont été déterminés par un calcul de fréquence sur une collection d'articles arabes de l'Agence France Press (AFP). Cette liste regroupe les préfixes et les suffixes les plus utilisés dans la langue arabe tel que les conjonctions, préfixes verbaux, pronoms possessifs, pronoms compléments du nom ou suffixes verbaux exprimant le pluriel etc…

Préfixes							
لا	في	لل	كم	بم	وت	بت	وال
با	وا	ليـ	فم	لم	ست	يت	فال
	فا	ويـ	ال	وم	نت	مت	بال
Suffixes							
ا	ة	ين	ية	هم	ته	وه	ات
	ه	يه	تك	هن	تم	ان	وا
	ي	ية	نا	ها	كم	تي	ون

Tableau 1 : Liste des préfixes et suffixes les plus fréquents

6. ***Calcul de fréquence*** calcule le nombre d'occurrences d'un mot significatif dans un document.

7. ***Expressions indicatives*** augmentent le poids des phrases qui pourraient apporter une information intéressante. C'est en quelque sorte le rôle inverse des mots vides. Le principe est d'utiliser un dictionnaire qui contiendrait l'ensemble de ces expressions. Il regroupe les annonces thématiques, les soulignements etc.

8. ***Calcul du poids des phrases*** par combinaison des différentes méthodes grâce à l'équation suivante [8], [9], [10]:

$$Sc = \alpha_1 Sc_{lead} + \alpha_2 Sc_{title} + \alpha_3 Sc_{cue} + \alpha_4 Sc_{tf \cdot idf}$$

$Sc_{lead} = 2$ pour la première phrase et $Sc_{lead} = 1$ pour les autres phrases.

$$Sc_{title} = \sum_{w \in S} a(w).tf(w)$$

$a(w) = 2$ si le mot w apparaît dans le titre de l'article et 0 dans les autres cas.

$tf(w)$ est la fréquence de w dans la phrase.

$$Sc_{cue} = \sum_{w \in S} c(w).tf(w)$$

$c(w) = 1$ quand w apparaît dans la liste des mots indicative sinon $c(w) = 0$.

$$Sc_{tf \cdot idf} = \frac{1}{|S|} \sum_{w \in S} \frac{tf(w) - 1}{tf(w)} \log \frac{DN}{df(w)}$$

$tf(w)$ est la fréquence de w dans la phrase S.

DN est le nombre total des documents dans le corpus.

$df(w)$ est le nombre de documents dans lesquels où w apparaît.

$|S|$ est le nombre de mots dans la phrase S.

Les coefficients α_i sont initialement fixés à 1. Ils peuvent être fixés arbitrairement ou déterminés de manière expérimentale (par apprentissage par exemple). L'avantage de cette technique est qu'elle permet d'ajuster ces coefficients suivant la nature des corpus. Le résultat sera retourné sous forme de liste de phrases triée par score.

9. **Extraction des phrases** permet de retourner le résultat final suivant le choix du pourcentage de compression. Ce pourcentage représente le nombre de phrases extraites par rapport au nombre de phrases contenues dans le document. Ces étapes ont été suffisantes pour extraire des résumés courts, le Tableau 2 donne une idée sur le nombre de mots et le taux de compression généré par Lakhas.

DOCSET	Source	Résumé	Résumé très court	Résumé/Source	Résumé très court/Source	Traduction de Ajeeb	Traduction de ISI
D1001	59	26	14	44%	23%	20	16
D1003	130	29	12	22%	9%	18	15
D1005	180	29	15	16%	8%	24	22
D1011	297	29	17	10%	6%	28	24
D1012	207	32	13	15%	6%	18	14
D1014	180	26	11	15%	6%	17	13
D1016	167	34	18	20%	11%	27	24
D1018	149	31	13	21%	9%	19	16
D1019	294	27	15	9%	5%	24	19
D1023	239	29	15	12%	6%	22	18
D1038	100	29	14	29%	13%	20	19
D1043	201	28	13	14%	6%	20	16
D30002	146	32	16	22%	11%	21	18
D30003	174	30	17	17%	10%	28	22
D30033	223	36	17	16%	8%	25	20
D30040	291	27	15	9%	5%	20	16
D30042	177	30	14	17%	8%	18	17
D30053	194	27	15	14%	8%	22	18
D31001	222	31	13	14%	6%	21	16
D31009	197	27	16	14%	8%	24	20
D31016	90	27	16	30%	18%	25	19
D31022	175	30	15	17%	9%	20	20
D31029	148	31	14	21%	10%	22	17
D31043	256	29	12	11%	5%	18	17
Moyenne	**187**	**29**	**15**	**16%**	**8%**	**21**	**18**

Tableau 2 : Moyenne du nombre de mots arabes par docset pour texte source et résumés, les deux dernières colonnes donnent le nombre de mots dans les textes anglais correspondants. Un docset représente un dossier de 10 documents fourni par le NIST.

Comme nous pouvons le voir dans le Tableau 2 colonne 3, les résumés produits par l'extraction ont environ 29 mots arabes en moyenne, mais l'évaluation à DUC a imposé que les résumés aient 10 mots anglais. Pour satisfaire cette contrainte et raccourcir nos résumés, nous avons dû développer des procédures spécifiques de compression, décrites dans la section suivante, pour satisfaire cette contrainte

3 Extension du système

Afin de réduire le nombre de mots dans les phrases sélectionnées comme résumé, nous avons appliqué quatre sortes de réductions.

Substitution de nom par suppression du poste ou la fonction par exemple pour *le secrétaire général des Nations Unis Kofi Annan* one ne garde que *Kofi Annan*

Suppression de mots non expressifs qui n'ajoutent pas d'information substantielle tels que des jours de la semaine ou les mois, les chiffres écrits en lettres, les adverbes, quelques conjonctions de subordination, etc.

Suppression de parties de phrases à partir de frontières comme la ponctuation, les conjonctions de coordination ou de subordination et dans certains cas des mots connecteurs

Suppression des constructions de discours indirect en ne gardant que le reste de la phrase en utilisant les modèles du Tableau 3.

Motif en arabe	Motif en français
أعلن X ... ان **R**	X a déclaré .... que **R**.
ذكر X ... ان **R**	X a mentionné ... que **R**
أفاد X ... ان **R**	X a reporté ... que **R**
أوضح X ... بان **R**	X a clarifié ... que **R**
أعلن X ... انه **R**	X a déclaré .... que cela est **R**.

Tableau 3 : Quelques modèles de motifs avec leur traduction en français

4 Évaluation

L'évaluation de résumés produits par des systèmes de résumé de texte automatiques est un processus complexe. La tâche d'évaluation est normalement exécutée manuellement par des juges qui comparent subjectivement des résumés différents et choisissent le meilleur. Le problème de cette approche est le fait que les juges qui exécutent la tâche d'évaluation ont souvent des idées très différentes sur ce qu'un bon résumé devrait contenir. Un autre problème avec l'évaluation manuelle est son coût en temps.

Néanmoins nous avons pu comparer la qualité de nos résultats au moyen de deux évaluations :

1- En comparant les résumés produits par Lakhas avec d'autres technologies de production de résumé automatique de texte arabe.
2- En participant à une compétition d'évaluation de résumé automatique sur des textes traduits à partir de l'arabe.

4.1 Comparaisons avec des systèmes commerciaux

Nous avons comparé nos résultats avec deux systèmes: Sakhr Arabic Summarizer[3] de la société Sakhr Software considérée comme un leader dans le traitement automatique de la langue Arabe et Pertinence Summarizer[4] de la société Pertinence Mining qui lit et condense des textes en 14 langues, dont l'arabe.

[3] http://www.sakhr.com

[4] http://www.pertinence.net

Notre expérimentation a porté sur un échantillon de 26 documents comportant phrases varie entre 4 à 13 phrases par document avec une moyenne globale de 6 phrases.

Taux de compression		Lakhas/Pertinence		Lakhas/Sakhr		Pertinence/Sakhr	
17%	1 phrase pertinente	19	73%	20	77%	17	65%
33%	2 phrases pertinentes	36	69%	30	58%	31	60%
50%	3 phrases pertinentes	57	73%	39	50%	37	47%

Tableau 4: Corrélations des systèmes Lakhas/Pertinence/Sakhr, les colonnes représentent le nombre total de phrases communes dans les résumés et leur pourcentage entre chaque paire de systèmes pour les 26 documents.

Le Tableau 4 indique que pour une compression de 17%, c'est-à-dire en résumant le texte en une phrase, nous avons obtenu pour les 26 documents, les 19 mêmes résumés que ceux de Pertinence et 20 pour ceux de Sakhr. Et pour une compression de 33% à deux phrases, nous avons obtenu 10 résumés identiques à 100% que ceux de Pertinence et les 16 autres recouvraient 50% des autres résumés de Pertinence.

Cette première expérimentation nous permet d'affirmer que Lakhas est compétitif avec les systèmes commerciaux, tous basés sur des techniques d'extraction. Lakhas donne des résultats satisfaisants du point de vue informationnel. Il reste encore à augmenter la taille de l'échantillon et à tester avec d'autres types de documents.

4.2 Évaluation à la compétition DUC 2004

Lakhas a été le premier système de résumé arabe à être formellement évalué et comparé avec des concurrents anglais dans une compétition internationale d'évaluation.

Afin de comparer nos résultats à ceux d'autres équipes à DUC, nous avons traduit nos résumés arabes par Ajeeb (http://english.ajeeb.com) un système de traduction arabe-anglais commercialisé sur le web.

NIST a évalué les résumés anglais avec ROUGE en utilisant 4 modèles de résumés comme références construits par des humains à partir des traductions manuelles. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) représente un système automatique d'évaluation de résumés. Ce système [11], dont les scores semblent bien corrélés avec l'évaluation humaine, inclut des mesures pour déterminer automatiquement la qualité d'un résumé en le comparant à d'autres résumés modèles créés par des humains. Les mesures comptent le nombre d'unités de recouvrement des n-grammes entre le résumé généré automatiquement et les résumés modèles, ces unités représentent les termes du texte pour lesquel on applique une lemmatisation par "Porter Stemmer".

Comme nous pouvons le voir sur le Tableau 5, les résultats Lakhas (LKS) sont très bons comparés à d'autres systèmes bien que nous ayons suivi une voie totalement différente. Regardant les résultats, nous avons conjecturé que les erreurs de traduction étaient principalement responsables de certains de nos points relativement mauvais.

ID	ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-4		ROUGE-L		R-W-1.2	
model	*0.395*		*0.147*		*0.064*		*0.027*		*0.344*		*0.200*	
LKS-ISI	**0.297**	**1**	**0.084**	**1**	**0.029**	**1**	**0.009**	**3**	**0.256**	**1**	**0.153**	**1**
142	0.218	7	0.076	2	0.029	2	0.010	1	0.201	6	0.126	4
134	0.259	2	0.047	10	0.011	13	0.002	16	0.220	2	0.129	2
LKS	***0.236***	***6***	***0.052***	***7***	***0.016***	***9***	***0.003***	***9***	***0.207***	***3***	***0.125***	***6***
8	0.255	4	0.075	3	0.026	3	0.009	2	0.207	4	0.127	3
59	0.255	3	0.071	4	0.023	4	0.006	4	0.206	5	0.126	5
...												
3	0.137	25	0.029	21	0.009	18	0.002	15	0.116	25	0.074	25

Tableau 5 : Score de Rouge pour quelques systèmes ainsi que leurs rang où LKS représente les résultats obtenus à partir des traductions avec Ajeeb.

Après la compétition grâce à la collaboration de Franz Och à ISI, nous avons obtenu une traduction anglaise de nos résumés arabes avec le système ISI, un des deux qui avaient été utilisés pour traduire les document originaux. Nos scores sont alors devenus les meilleurs de tous (voir Tableau 5) où la ligne de LKS-ISI représente les nouvelles valeurs.

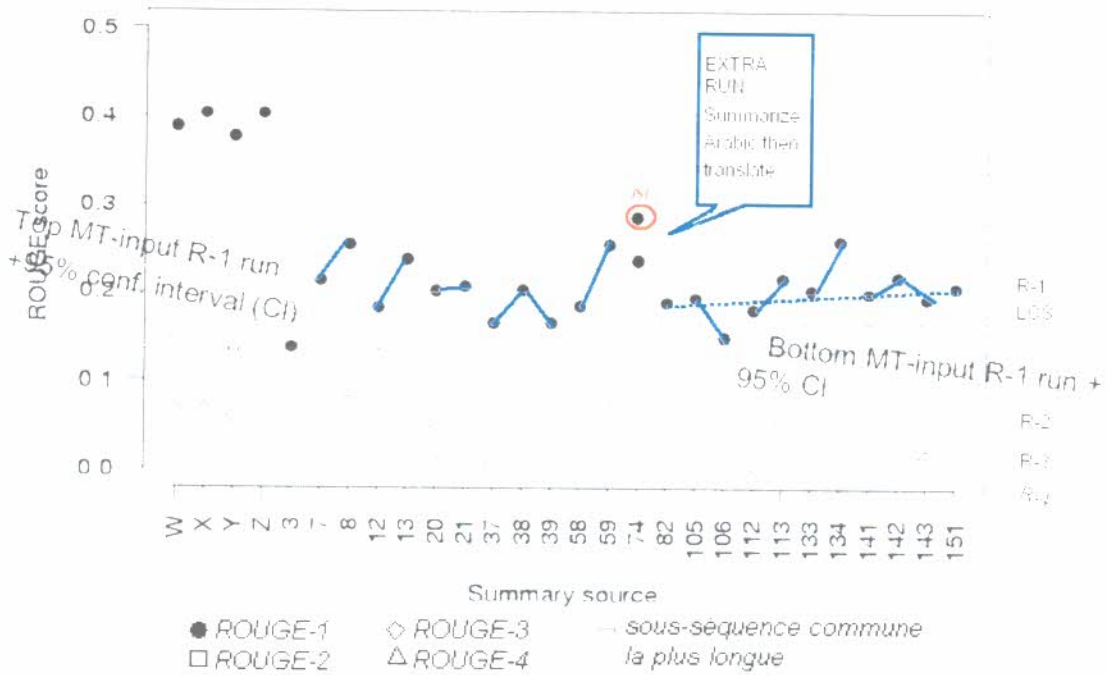

Figure 2 : Score de Rouge par participant, présenté par Paul Over à DUC 2004 [12], les abscisses représentent les participants et les modèles W, X, Y, Z, et les différent scores de ROUGE en ordonnées

De même que sur la Figure 2, on peut voir le nouveau score ROUGE-1 obtenu sur les résumés traduis par ISI est impressionnant, même dans le cas où les autres concurrents ont utilisé les traductions manuelles.

Les procédures de réduction que nous avons ajoutées à Lakhas, en particulier la suppression des constructions de discours indirect, ont permis à Lakhas de se distinguer des autres participants par la préservation de plus d'information pertinente.

Bien que les traductions faites par Ajeeb paraissaient appropriées, les résultats de LKS-ISI par rapport à LKS étaient beaucoup meilleurs pour les deux raisons suivantes:

- Les traductions d'ISI produisent une moyenne de 3.5 mots de moins que Ajeeb, et ignorent les mots inconnus, alors que Ajeeb les garde et essaye d'insérer un mot par décomposition ou translittération, ce qui souvent produit des mots sans valeur, la 7ème et 8ème colonne du Tableau 2 montre la différence en nombre de mots entre les deux systèmes.

- Les mots traduits par ISI sont habituellement identiques aux mots utilisés dans les modèles de référence tandis que ceux d'Ajeeb sont souvent synonymes

Model	Traduction de Ajeeb Traduction de ISI	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	R-W-1.2
King Hussein nearly **finished** *chemotherapy* **treatments** at American Mayo Clinic	Al-Malik Hussain ended the fourth stage from a *chemotherapy* from origin six|stages	0.13	0.03	0.00	0.00	0.13	0.09
	King Hussein finished the fourth phase of the chemical **treatment** of the six|stages	0.42	0.15	0.03	0.00	0.39	0.22
Nelson Mandela arrives to participate in **annual** Gulf States **summit**	Nelson Mandela arrived to the United Arab Emirates for the participation in| the annual summit to the Gulf countries	0.35	0.11	0.06	0.04	0.33	0.20
	Nelson Mandela arrived in the Emirates to participate in the **annual summit**| of the Gulf	0.60	0.28	0.16	0.00	0.55	0.32
Cohen confident Gulf *countries* will support **"appropriate action" against Iraq**	The Gulf Arab *countries* will offer the support to the doing of a suitable w|ork against Iraq	0.37	0.06	0.00	0.00	0.34	0.21
	Gulf Arab states will support for **"appropriate action" against Iraq,**	0.55	0.27	0.13	0.04	0.53	0.32

Tableau 6 : Les scores de ROUGE pour des traductions de phrases par Ajeeb et ISI. Les mots soulignés n'ont pas été pris en compte pendant l'évaluation en raison de la troncature. Italique/Gras est pour des mots trouvés dans la traduction d'Ajeeb/ISI et également dans le modèle

Comme on peut voir dans le Tableau 6, en plus de la fréquence du même mot entre ISI et le modèle, les bi-grammes sont aussi plus présents dans ISI même quand les mots existent dans Ajeeb (exemple *to_participate*, *will_support*, ...).

ROUGE semble être un outil très intéressant pour l'évaluation de résumés, mais il dépend des modèles de référence utilisés.

5 Conclusion

Les méthodes d'extraction se sont avérées adaptables à l'arabe et nous avons pu faire nos expérimentations sur le corpus Gigaword qui regroupe un ensemble de nouvelles journalistiques en arabe.

Notre participation à DUC 2004 en suivant une autre approche que celle proposée par le NIST nous a permis d'évaluer notre système. Cette expérimentation est intéressante par son approche et ses résultats, elle a montré qu'il est plus pratique de traiter des données qui respectent en général une certaine structure ; de plus, le travail sur les textes en langue originale donne accès à plus d'informations fiables.

Dans les traitements automatiques de la langue où l'ambiguïté est omniprésente que ça soit en traduction, en recherche d'information ou en résumé automatique, faisant appel à des techniques basées sur la statistique, la linguistique et les règles, la moindre omission d'informations clés influence négativement les résultats.

Les résultats obtenus sont très bons mais leurs évaluations par ROUGE semblent influencées par les systèmes de traductions, ceci est dû au fait que ROUGE utilise comme référence des modèles générés par des humains qui ont travaillé sur les textes traduits manuellement.

Malheureusement, il n'existe pas jusqu'à ce jour de méthode de validation, sauf le fait de participer à des compétitions d'évaluation ou de disposer de ressources nécessaires comme des données de jugements pour calculer les précision et rappel.

Remerciements

Nous tenons à remercier Paul Over du NIST pour sa collaboration et son aide pour l'obtention des correspondances pour les textes arabes et également Franz Och d'ISI pour nous avoir fourni les traductions en anglais de nos résumés arabes en utilisant le système de traduction automatique d'ISI.

Références

[1] C. Jaruskulchai and C. Kruengkrai: A Practical Text Summarizer by Paragraph Extraction for Thai. *The Sixth International Workshop on Information Retrieval with Asian Languages (IRAL-2003), July 7, 2003, Sapporo Japan, pp. 9-16.*

[2] Larkey L. S., Ballesteros L. and Connell M., Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis, *In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 2002, pp. 275-282.*

[3] M. Attia, A large-scale computational processor of the Arabic morphology, *A Master's Thesis, Cairo University, (Egypt) 2000.*

[4] Mohammed Aljlayl and Ophir Frieder. On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach, *In 11th International Conference on Information and Knowledge Management (CIKM)*, November 2002, *pp. 340-347.*

[5] Darwish, K. and D. Oard. CLIR Experiments at Maryland for TREC 2002: Evidence Combination for Arabic-English Retrieval *in TREC. 2002. Gaithersburg, MD.* http://trec.nist.gov/pubs/trec11/papers/umd.darwish.pdf

[6] K. Darwish: Probabilistic Methods for Searching OCR-Degraded Arabic Text, *Doctoral dissertation, University of Maryland, 2003*

[7] A. Chen and F. Gey: Building an Arabic Stemmer for Information Retrieval. *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*. National Institute of Standards and Technology, Nov 18-22, 2002, *pp. 631-640*.

[8] H. Saggion, Génération automatique de résumés par analyse sélective, *Thèse de Ph.D en Informatique, Université de Montréal, août 2000.*

[9] C. Nobata & S. Sekine, Results of CRL/NYU System at DUC-2003 and an Experiment on Division of Document, *Proceedings of the Document Understanding Conference (DUC 2003), Edmonton, Canada. pp. 79-84*

[10] Ishikawa, K., Ando, S., Okumura, A.: Hybrid Text Summarization Method based on the TF Method and the Lead Method. *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization. Tokyo. Japan. March 2001. pp. 219-224.*

[11] C-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, pp. 74-81.*

[12] P. Over, J. Yen, An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems, *Proceedings of the Document Understanding Conference (DUC 2004), Boston (USA), May 6-7 2004, pp. 1-21.*