Une approche hybride de résumé des textes arabes

AIDJOULI Hocine
Université Djelfa

Introduction

Cet article est consacré à la description de notre approche - proposé pour le résumé automatique des textes en langue arabe – basé fondamentalement sur les méthodes d'extractions de phrases.

Nous commençons d'abord par une brève synthèse d'une partie théorique à savoir : le traitement de l'arabe dans la grammaire HPSG et les méthodes de résumé automatique puis nous présenterons notre conception ainsi que l'architecture de notre approche et nous détaillerons chacun des modules le composant.

1. Présentation – synthèse

Rappelons que l'objectif des grammaires HPSG consiste à réduire le nombre de règles syntaxiques grâce au grand nombre d'informations présentes dans les représentations lexicales [Qsi 09]. En effet, le formalisme HPSG permet de représenter, dans les entrées lexicales, des informations d'ordre syntaxique ou sémantique. C'est sans doute grâce à cette richesse de représentation que HPSG est adoptée par plusieurs chercheurs en informatique linguistique.

Aussi, rappelons que la langue arabe possède, outre les caractéristiques de base partagées avec d'autres langues comme le latin, des caractéristiques très particulières comme, l'absence de voyelles et de ponctuations régulières, la flexion et l'agglutination [Din 05]. Ainsi, le formalisme à choisir pour effectuer l'analyse syntaxique de l'Arabe doit avoir la capacité de supporter facilement des extensions ou des adaptations pour prendre en charge les dites caractéristiques spécifiques à l'arabe.

En effet, pour tenir en compte aussi bien des caractéristiques de bases que celles spécifiques à l'Arabe, lors de l'analyse, et afin de décrire les constructions langagières à l'aide d'un petit nombre d'opérateurs, nous avons opté pour l'application des grammaires HPSG. Notre choix se justifie par le fait que HPSG permet de représenter, dans les entrées lexicales, des informations d'ordre phonologique, morphologique, syntaxique, sémantique et pragmatique. Cette richesse nous permettra d'éviter les solutions parasites dues, par exemple, à la non voyellation et à l'agglutination par l'exploitation des traits syntaxiques et sémantiques des matrices attribut/valeur des mots. Ces traits donnent des indices sur les types des mots attendus dans la phrase sans pour autant avoir recours aux règles de réécriture [Cho 95].

De plus, HPSG repose sur le cadre formel de la logique attribut/valeur. Les propriétés formelles sont ainsi décrites dans un cadre homogène utile lors de l'implémentation.

Par ailleurs, HPSG permet de tenir en compte un maximum de phénomènes linguistiques et de décrire les constructions langagières à l'aide d'un nombre réduit d'opérateurs.

Rappelons aussi, comme nous l'avons présenté au premier chapitre, qu'il existe quelques techniques pour le résumé automatique de textes :

les méthodes d'extraction, les méthodes hybrides et les méthodes de compréhension et génération.

A. <u>Les méthodes d'extraction de phrases :</u>

L'objectif des méthodes d'extraction des phrases est de repérer dans le texte source les phrases les plus importantes. Le résultat obtenu est alors un extrait du texte source [Mon 04].

A.1 Méthodes à base de mots clés

Cette méthode est basée sur le fait que l'auteur se sert (pour exprimer ses idées principales) de quelques mots-clés qui ont tendance à être récurrents dans le texte [Par 02]. Le résumé automatique est alors produit en recherchant dans le texte source les unités de texte minimales réunissant ses mots-clés. Ce principe est souvent appliqué en différentes variantes :

A.1.1 Mots-clé prédéfinis

La liste de mots-clés peut être introduite par l'utilisateur (domaine d'intérêt) ou composée des mots-clés établis par l'auteur.

A.1.2 Titres

Étant donné que le titre est l'expression la plus significative et qui résume le mieux un document en quelques mots, on peut dire que la phrase qui ressemble le plus au titre est la plus marquante du document. Par conséquent, on peut attribuer à chaque phrase un poids en fonction de sa ressemblance avec le titre [Ish 01].

Dans ce cas on considère les mots du titre du texte comme des mots-clés et on produit le résumé en sélectionnant les phrases qui couvrent certains mots apparaissant dans un titre .

A.1.3 Méthode de distribution de termes

L'idée de cette méthode est de considérer comme « importantes » les phrases qui contiennent des mots « importants » du texte. Un mot est considéré important s'il est employé assez fréquemment dans le texte [Luh 58].

A.2 La méthode de la position

La méthode de la position considère que les premières et dernières phrases de chaque paragraphe sont importantes. La méthode considère aussi des phrases positionnées dans certaines sections conceptuelles importantes, par exemple dans «Introduction» et «Conclusion» [Edm 69].

L'inconvénient de cette méthode est qu'elle dépend de la nature du texte à résumer ainsi que du style de l'auteur.

A.3 Expressions indicatives

Dans la littérature scientifique, on trouve souvent des expressions qui, indépendamment du domaine particulier du texte, font référence à des catégories conceptuelles. Ainsi, si dans l'introduction d'un article on trouve une phrase avec l'expression «L'objectif de cet article est...», on peut être presque sûr que ce qui suit dans la phrase est l'information sur les objectifs de l'article et lorsqu'on trouve une phrase qui commence par «Pour en conclure...», on a la certitude qu'il s'agit d'une conclusion.

L'idée de la méthode des expressions indicatives est de sélectionner dans tout le texte des phrases contenant ces types d'expression [Pai 81].

A.4 Cohésion lexicale

Construire des relations de cohésion lexicale entre les phrases et classifier les phrases selon «début de thème», «continuation de thème», «clôture de thème» et «marginale».

Sélectionner un sous-ensemble de phrases qui introduisent, continuent et terminent les thèmes [Ben 95].

A.5 *Chaînes lexicales*

Construire des chaînes lexicales qui lient des phrases contenant des mots liés par des relations de cohésion lexicale [Bar 97].

Sélectionner un sous-ensemble de chaînes et ensuite un ensemble de phrases de chaque chaîne.

A.6 *Extraction des paragraphes*

Calculer les connections entre les paragraphes d'un texte en utilisant des mesures de similarité [Sal 97].

Sélectionner les paragraphes les plus connectés.

A.5 *Classification des éléments*

Dans les textes de science et technique il y a des phrases qui font référence à des catégories conceptuelles telles que : Connaissances Antérieures, Contenu, Méthode et Résultat, on peut également constater que dans les résumés de science et technique des informations relatives à ces catégories sont souvent retenues pour le résumé [Lar 02].

B. *Approches hybrides*

Les méthodes présentées dans les sections précédentes utilisent des traits (fréquence, position, expression indicative, etc.) qui ne peuvent isolément garantir des résultats optimaux.

On combine souvent ces traits pour obtenir des meilleurs résultats [Str 98].

C. *Méthodes de compréhension et génération*

A la différence des mesures «quantitatives» attribuant un poids à chaque phrase, les méthodes de compréhension et génération essaient de découvrir comment chaque phrase contribue à l'organisation du texte, quelle est la fonction de chaque phrase dans le tout.

Pour produire un bon résumé il n'est pas suffisant de repérer l'information importante mais aussi de la régénérer.

L'inconvénient de ces méthodes est que sa mise en œuvre reste extrêmement difficile.

Les méthodes d'extraction offrent certains avantages :

- Simplicité de mise en œuvre,
- Rapidité de traitement,
- Indépendance des traitements par rapport à la langue,
- Une compression paramétrable en modifiant les seuils de sélection.

Une méthode seule ne peut donner de bons résultats, l'approche mixte (hybride) qui combine plusieurs méthodes d'extraction est souvent utilisée [Hor 00].

Pour cela nous proposons une méthode que nous avons baptisé AWDJIZ (résume en français) est une approche d'extraction de résumé automatique de texte en langue arabe basée essentiellement sur deux idées fédératrices :

- L'usage de la grammaire HPSG adapté à la langue arabe pour identifier et représenter les objets linguistiques,
- L'utilisation d'un approche hybride d'extraction qui a prouvé son efficacité pour d'autres langues et semble donner des résultats satisfaisants à la langue arabe.

La mise en oeuvre fonctionnelle de AWDJIZ est représentée à la figure 01. Elle repose sur l'identification des objets linguistiques par la grammaire HPSG ainsi que sur la combinaison de quelques méthodes afin de permettre la génération de résumé.

AWDJIZ est composé en deux phases :

Phase 01 : Analyse en HPSG du texte introduit

Phase 02 : Application d'une méthodes hybride d'extraction

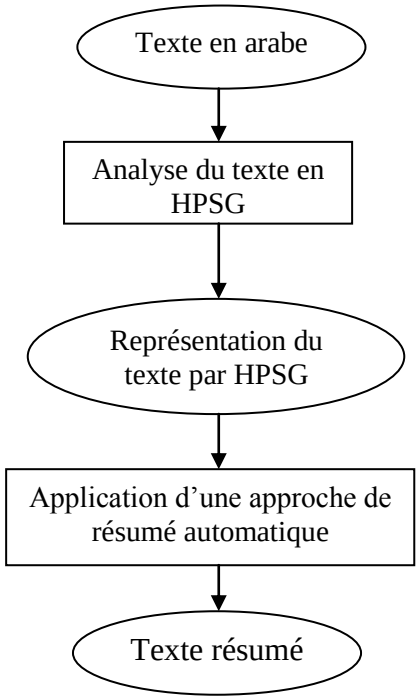

Figure 01 : Schéma du traitement

3. Schéma général de AWDJIZ

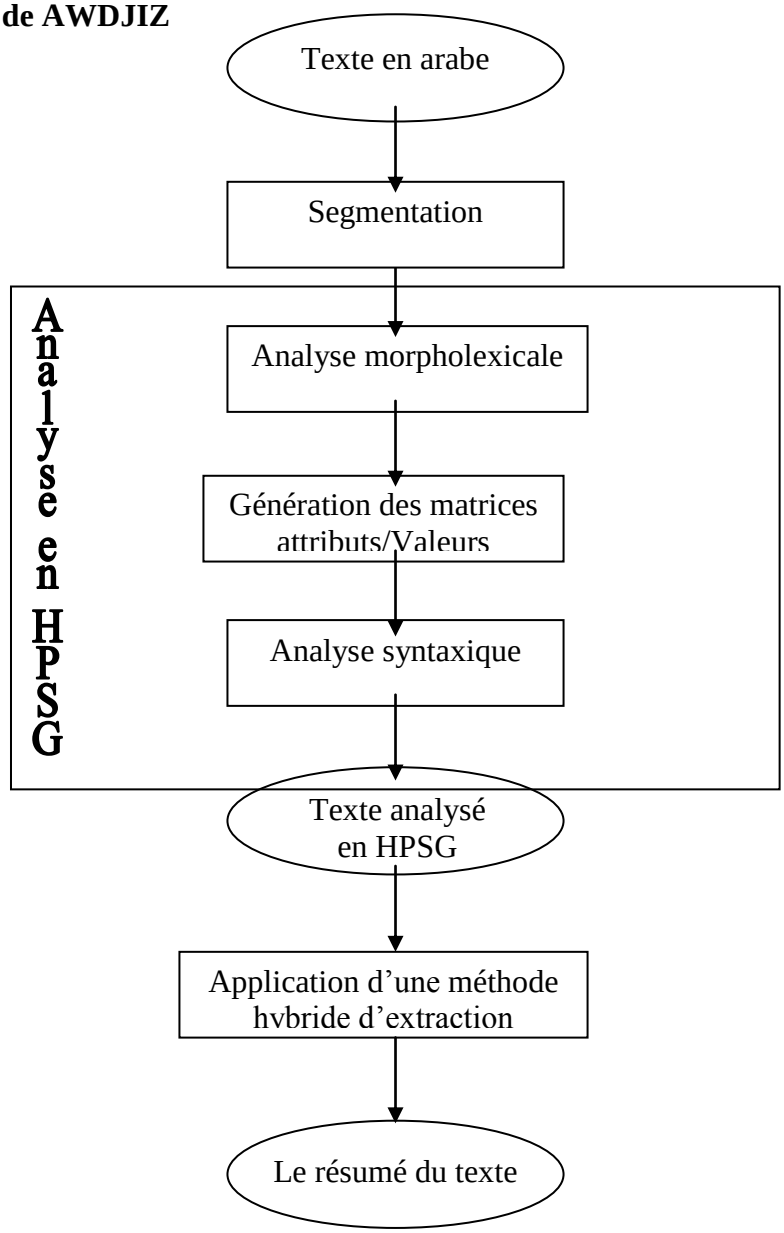

Figure 02 : Schéma global de AWDJIZ

4. Description des principaux modules composant AWDJIZ

4.1 Segmentation et détection de la racine:

La segmentation consiste à détecter les phrases, les mots et les racines.

Pour la segmentation de texte en phrases, Riadh OUERSIGHNI [Oue 01] utilise à la fois :

- Une segmentation morphologique basée sur la ponctuation,
- Une segmentation basée sur la reconnaissance de marqueurs morphosyntaxiques ou des mots fonctionnels comme :أو ، و ، أي ، لكن ، حتى، ou, et, c.a.d., mais, quand.

Cependant, ces derniers particules peuvent jouer un autre rôle que celui de séparer les phrases.

Dans notre cas, nous identifions les mots en effectuant une segmentation qui se base sur les indicateurs de surfaces (ou d'espaces).

Finalement on utilise le tableau suivant pour éliminer les préfixes et les suffixes qui peuvent être collé à la racine (les racines en arabe contiennent généralement 3 à 4 caractères).

Nous utilisons la liste de préfixes et de suffixes proposé par K. Darwish [Dar 03] . Plusieurs d'entre eux ont été utilisés par A. Chen and F. [Che 02] pour la lemmatisation de mots arabes :

Préfixes							
لا	فبـ	للـ	كمـ	بمـ	وتـ	بتـ	والـ
با	وا	ليـ	فمـ	لمـ	ستـ	يتـ	فالـ
	فا	ويـ	الـ	ومـ	نتـ	متـ	بالـ
Suffixes							
ا	ة	ين	ية	هم	ته	وه	ات
	ه	يه	تك	هن	تم	ان	وا
	ي	ية	نا	ها	كم	تي	ون

Listes des préfixes et suffixes les plus fréquents

4.2 Analyse morpho-lexicale :

Cette étape prend en entrée les résultats de la segmentation et fournit en sortie un lemme accompagné de propriétés morphologique, syntaxiques et sémantiques.

L'utilisation de la grammaire HPSG commence par cette étape, Mourad Loukam [Lou 08] a donné une grande importance pour connaître les caractéristiques de chaque item.

4.3 Génération des matrices attributs/Valeurs :

Consiste à générer pour chaque item sa structure de trait sous la forme d'une matrice attributs/valeurs (AVM).

4.4 Analyse syntaxique :

Cette analyse détermine si une phrase (ou une succession de mots) appartient ou non au langage et respecte donc les règles de la grammaire de la langue arabe [Fer 03].

L'analyse syntaxique en HPSG se base principalement sur l'application du processus d'unification. Il opère sur les structures de traits (AVM) des entrées lexicales des différents mots, déjà générées lors de la phase précédente, ainsi que sur les règles syntaxiques (schémas).

Ce module inspiré du projet SYNTAXE des chercheurs Noureddine Loukil, Kais Haddar et Abdelmajid Ben Hamadou [Ham 09] , ce projet permet l'analyse syntagmatique pour la reconnaissance des syntagmes nominaux, verbaux et prépositionnels, et produit des arborescences des phrases si elles sont totalement acceptées par la grammaire ou des arborescences partielles des syntagmes reconnus par l'analyseur.

4.5 Résultat du traitement du texte en HPSG :

Il s'agit de présenter sous forme concrète (AVM) la représentation syntaxique et sémantique du texte analysé.

4.6 Application d'une méthode hybride d'extraction :

Consiste à choisir parmi les méthodes citées en haut celles qui donnent les meilleures résultats à savoir : les méthodes de mots clés, titre, position, expression indicative.

Si le texte est de type scientifique ou technique on peut ajouter la méthode de classification des éléments.

Méthode $_{\text{hybride}}$ = Méthode $_{\text{mots clés}}$ + Méthode $_{\text{titre}}$ + Méthode $_{\text{position}}$ + Méthode $_{\text{expression indicative}}$.

Le résultat est l'union des phrases obtenues par chacune des méthodes appliquées par ordre d'apparition de ces phrases dans le texte original et sans répétition (par exemple si une phrase est donnée par deux méthodes, elle ne sera prise qu'une seule fois en résumé).

Prenons des exemples :

Pour la méthode mots clés : Cette méthode est basée sur le fait que l'auteur se sert (pour exprimer ses idées principales) de quelques mots-clés qui ont tendance à être récurrents dans le texte. Le résumé automatique est alors produit en recherchant dans le texte source les unités de texte minimales réunissant ses mots-clés.

Ex. : intelligence artificielle, traitement automatique de la langue, résumé automatique

Pour la méthode titre : Étant donné que le titre est l'expression la plus significative et qui résume le mieux un document en quelques mots, on peut dire que la phrase qui ressemble le plus au titre est la plus marquante du document.
Dans ce cas on considère les mots du titre du texte comme des mots-clés et on produit le résumé en sélectionnant les phrases qui couvrent certains mots apparaissant dans un titre.
Ex. Le résumé automatique
Pour la méthode position : La méthode de la position considère que les premières et dernières phrases de chaque paragraphe sont importantes car elles sont considérés comme thématiques, c'est-à-dire elles «résument» le contenu du paragraphe. La méthode considère aussi des phrases positionnées dans certaines sections conceptuelles importantes, par exemple dans «Introduction» et «Conclusion».
Pour la méthode expression indicative : dans la littérature scientifique, on trouve souvent des expressions qui, indépendamment du domaine particulier du texte, font référence à des catégories conceptuelles. Ainsi, si dans l'introduction d'un article on trouve une phrase avec l'expression «L'objectif de cet article est...», on peut être presque sûr que ce qui suit dans la phrase est l'information sur les objectifs de l'article et lorsqu'on trouve une phrase qui commence par «Pour en conclure...», on a la certitude qu'il s'agit d'une conclusion.
L'idée de la méthode des expressions indicatives est de sélectionner dans tout le texte des phrases contenant ces types d'expression.
Ex. finalement, voila comme résultats, …….
Finalement pour la méthode de classification des éléments : Dans les textes de science et technique il y a des phrases qui font référence à des catégories conceptuelles telles que : Connaissances Antérieures, Contenu, Méthode et Résultat, on peut également constater que dans les résumés de science et technique des informations relatives à ces catégories sont souvent retenues pour le résumé.
Ex. les résultats obtenus sont les suivants : ……
Si on combine les résultats de chaque méthode, par ordre d'apparition des phrases dans le document source et sans répétition, on arrive bien à un très bon extrait (significatif) du texte original.

4.7 Extraction des phrases :

Permet de retourner le résultat final selon le choix du pourcentage de compression (choix de l'utilisateur). Ce pourcentage représente le nombre de phrases extraites par rapport au nombre de phrases contenues dans le document.

5. Conclusion

Dans cet article, nous avons décrit notre approche utilisée pour la production automatique de résumés en arabe, en décrivant l'architecture de AWDJIZ, qui se base sur une phase préliminaire concernant l'analyse du texte par la grammaire syntagmatique guidée par les têtes HPSG et une autre phase qui consiste à l'application d'une méthodes hybride d'extraction (qui combine plusieurs méthodes d'extraction afin de donner de meilleurs résultats).

Cette approche ce voit puissante vu :

- L'analyse des textes arabes par la grammaire syntagmatique guidée par les têtes (HPSG), le model le plus utilisé au monde dans la majorité des applications informatiques.
- La génération de résumé se fait par extraction des phrases pertinentes qui sont obtenues par une méthode hybride formée par la combinaison de plusieurs méthodes
- Sa simplicité en terme d'implémentation

Bibliographie :

[Bar 97] : Barzilay and Elhadad, Using Lexical Chains for Text Summarization. In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 10-17, Madrid, Spain 1997.

[Ben 95] : Benbrahim and Ahmad, Text Summarisation : the Role of Lexical Cohesion Analysis. The New Review of Document & Text Management, pages321-335, 1995.

[Che 02] : 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, pp631-640.

[Cho 65] : Chomsky Naom, Aspects of the theory of syntax. MIT Press France 1965

[Dar 03] : K. Darwish, Building a Shallow Arabic Morphological Analyzer in One Day. *Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA. pp. 47-54.*

[Din 05] : Dina El Kassas, thèse de doctorat 'Une étude contrastive de l'arabe et du français dans une perspective de génération multilingue', Université Paris 7, 2005

[Dou 04] : F. S. Douzidia, Résumé automatique de texte arabe Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique Septembre, 2004, Université de Montréal, Canada, 2004.

[Edm 69] : H. P. Edmundson: New methods in automatic abstracting, *Journal of the Association for Computing Machinery (ACM), vol. 16 N°2 pp. 264-285, April 1969.*

[Fer 03] : Soufiane Ferfera, Proposition d'une architecture multi-agents pour l'analyse syntaxique de la langue arabe, mémoire de magister CRSTDLA Alger 2003

[Ham 09] : Noureddine Loukil, Kais Haddar, Abdelmajid Ben Hamadou Laboratoire de Recherche en Informatique et Multimédia, Sfax, Tunisie Normalisation de la représentation des lexiques syntaxiques arabes pour les formalismes d'unification www.miracl.rnu.tn 2009 Consulté le 31/03/2010

[Hor 00] : Horacio Saggion, Génération automatique de résumé par analyse sélective Thèse de doctorat, Université de Montréal 2000.

[Ish 01] : 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. pp.5-219-5-224, March 2001.*

[Lar 02] : Larkey, L. S., Ballesteros, L. and Connell M., improving Stemming for Arabia Information Retrieval: Light Stemming and co-occurrence Analysis, In proceeding of the 25th annual International conference on Research and development in information Retrieval (SIGIR 2002), Tampere, Finland, August 2002.

[Lou 08] : Mourad Loukam, une plate-forme d'analyse basée sur le formalisme HPSG pour l'Arabe standard, Université de Chlef 2008

[Luh 58] : P. H. Luhn, The Automatic Creation of Literature Abstracts, IBM Journal April 1958, pp. 159-165, April 1958.

[Mon 04] : M. Monod, Le résumé automatique, un petit état de l'art 2004

[Oue 01] : Riadh OUERSIGHNI, Modélisation des expressions figées en arabe en vue de la constitution d'une base de données lexicale. Thèse de doct., Univ. Lyon 2. 2001

[Par 02] : T.A.S. Pardo, L.H.M. Rino, M.G.V. Nunes, Extractive summarization: how to identify the gist of a text, *International.Bibliographie* 63 *Information Technology Symposium - I2TS 2002,*

Florianópolis-SC, Brazil, pp.245-260, 01-05 October 2002.

[Qsi 09] : Y. BAHOU L. HADRICH BELGUITH C. ALOULOU A. BEN HAMADOU Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés, Laboratoire de recherche LARIS Faculté des Sciences Economiques et de Gestion de Sfax – TUNISIE http://www.tsi.enst.fr/afrif/rfia2006/pdf/255.pdf 2009 Consulté le 31/03/2010

[Sal 97] : Salton, G., Singhal, A., Mitra, M., and Buckley, C., Automatic Text Structuring and summarization. Information Processing & Management, 33(2) :193-207, 1997.

[Str 98] : T. Strzalkowski, J. Wang and B. Wise, Summarization-based Query Expansion in Information Retrieval, *Proceedings of 36th Annual Meeting of the ACL, V. 2, pp. 1258-1264, Montreal 1998.*