Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules

Lamia Hadrich Belguith (1), Leila Baccour (1) et Ghassan Mourad (2)

(1) Laboratoire de recherche LARIS – Faculté des Sciences Economiques et de Gestion de Sfax
B.P. 1088, 3018, Sfax, Tunisie
l.belguith@fsegs.rnu.tn
leila_freind@techemail.com
(2) Equipe LaLICC – Paris Sorbonne
96, Bd Raspail, 75006 Paris
Ghassan.Mourad@paris4.sorbonne.fr

Mots clés : Segmenteur de textes arabes, segmentation en phrases, exploration contextuelle, expressions rationnelles.

Keywords: Arabic text tokenizer, sentence tokenization, contextual exploration, regular expressions.

Résumé Nous proposons dans cet article une approche de segmentation de textes arabes non voyellés basée sur une analyse contextuelle des signes de ponctuations et de certaines particules, tels que les conjonctions de coordination. Nous présentons ensuite notre système STAr, un segmenteur de textes arabes basé sur l'approche proposée. STAr accepte en entrée un texte arabe en format txt et génère en sortie un texte segmenté en paragraphes et en phrases.

Abstract We propose in this paper an approach to segment non-vowelled Arabic texts. Our approach is based on a contextual analysis of the punctuation marks and a list of particles, such as the coordination conjunctions. Then, we present our system STAr, a tokenizer based on the proposed approach. The STAr input is an Arabic text (in .txt format) and its output is a segmented text into paragraphs and sentences.

Introduction

Pour la plupart des applications de traitement automatique des langues naturelles (e.g., l'analyse de texte, l'extraction d'information, le résumé automatique) la segmentation devient une phase importante pour repérer les segments contenant les informations recherchées. Ainsi par exemple, commencer une analyse d'un texte sans le segmenter en phrases conduit à des résultats peu fiables; de même, avoir un mauvais segmenteur conduit à accumuler les erreurs du traitement automatique du texte (Mourad, 2001).

La segmentation consiste à désambiguïser les frontières des phrases et des paragraphes et se base généralement sur un ensemble de règles de segmentation. C'est une phase non triviale pour toute application en TALN. En effet, segmenter un texte nécessite le repérage des frontières formelles marquées par des signes typographiques. Par ailleurs, dans les textes arabes actuels, les signes de ponctuation ne sont pas très utilisés et dans le cas où ils y figurent, ils ne sont pas gérés par des règles d'utilisation. De plus, d'après l'observation de corpus, nous avons constaté que certaines particules (e.g., "و" (et), "ف" (donc)) jouent un rôle principal dans la séparation de phrases.

Dans ce qui suit, nous présentons un bref aperçu sur les travaux de segmentation. Ensuite, nous détaillons les difficultés rencontrées lors de la segmentation des textes arabes. Nous proposons, ensuite, notre approche de segmentation de textes arabes. Après, nous présentons le système STAr, un segmenteur de textes arabes basé sur l'approche proposée. Enfin, nous présentons l'évaluation de STAr.

1 Bref aperçu sur les travaux de segmentation

Les travaux sur la segmentation ne sont pas nombreux. Pour certaines langues latines, ils existent des segmenteurs fonctionnels. Alors que pour l'arabe, il y a peu de travaux sur la segmentation de textes en phrases et il n'existe pas des segmenteurs fonctionnels et spécifiques à l'arabe.

Dans ce qui suit nous présentons quelque segmenteurs pour le français et l'anglais.

- Le segmenteur INTEX (Silberztein, 93) utilise un transducteur pour découper un texte français en phrases en s'appuyant sur les signes de ponctuation.
- Le segmenteur SATZ (Palmer, Hearst, 1994) de textes anglais utilise les catégories lexicales au voisinage des signes de ponctuation et applique une méthode d'apprentissage en utilisant les réseaux de neurones.
- Le système SegATex (Mourad, 2001) est un segmenteur de textes français qui conçoit des règles de segmentation en étudiant les voisinages des signes de ponctuation et des marques typographiques en appliquant la méthode d'exploration contextuelle (DESCLES, 1997).

2 La segmentation de textes arabes : particularités et difficultés

La segmentation automatique de textes arabes présente plusieurs difficultés spécifiques à la langue arabe. Nous présentons dans ce qui suit certaines ambiguïtés qui rendent la segmentation difficile à réaliser sans une étude approfondie sur un corpus à large couverture.

- L'ambiguïté vocalique des mots : un texte arabe non voyellé est fortement ambigu. La proportion des mots ambigus passe à plus de 90% si les comptages portent sur les voyellations globales de ces mots (Debili, Achour, Souissi, 2002). Ainsi, un mot non voyellé peut avoir plusieurs caractéristiques morphologiques possibles (Chaâben, Belguith, 2003). Par exemple le mot "فهم" peut être un nom, un verbe, ou un pronom personnel précédé d'une conjonction de coordination.
- L'ambiguïté dérivationelle : le mot arabe n'est pas le résultat d'une simple concaténation de morphèmes comme c'est le cas pour l'anglais mais c'est à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un schème morphologique qu'on obtient un mot (Beesley, 1996). Ainsi, l'identification de la catégorie grammaticale de certains mots est ambiguë ce qui entraîne des difficultés au niveau de la segmentation automatique.
- L'ambiguïté structurelle : la phrase arabe est relativement longue et complexe en comparaison avec d'autres langues, tels que le français ou l'anglais. Ainsi il n'est pas rare de trouver des phrases arabes composées de plusieurs dizaines de mots.
- L'utilisation des signes de ponctuation : l'arabe n'est pas appuyée principalement sur les signes de ponctuations et les marqueurs typographiques; ces derniers ont généralement un rôle pausale. Ainsi, nous pouvons trouver tout un paragraphe arabe ne contenant aucun signe de ponctuation à part un point à la fin de ce paragraphe.
- L'agglutination : les conjonctions de coordinations jouent un rôle important dans la segmentation de textes arabes. Cependant, elles sont toujours agglutinées aux mots qui les suivent. Ainsi par exemple la lettre "و" (w) dans le mot "وهم" peut représenter une lettre du mot en question (i.e., « wahmun » (imagination)) ou une conjonction de coordination suivie d'un pronom personnel (i.e., "و"+ "هم" « wa+hum » (et + ils)).

3 Approche proposée pour la segmentation de textes arabes

Afin de surmonter les problèmes de segmentation que nous venons de présenter, nous proposons une approche de segmentation de textes arabes basée sur l'exploration contextuelle des signes de ponctuation, des mots connecteurs jouant le rôle de séparateur de phrases (e.g., "لكن" (lakin), "لقد" (laqad) et "أمّا" ('amma)) ainsi que celles de certaines particules tel que les conjonctions de coordination ("و" (wa) et "ف" (fā)).

L'exploration contextuelle repose sur une étude des indices linguistiques déclencheurs appelés indicateurs et des indices complémentaires associés à ces indicateurs et sur un ensemble de règles (Descles, 1991). Ainsi, nous proposons d'utiliser l'exploration contextuelle pour étudier les contextes droit et gauche de chaque mot ou particule jouant le rôle de séparateur de phrases. Pour ce faire, nous avons étudié un corpus de textes issus de quatre livres de l'enseignement tunisien (voir figure 1).

Corpus	Nombre de textes	Nombre de paragraphes	Nombre de mots
Livre de 5ème année primaire	70	618	19 254
Livre de 6ème année primaire	72	173	18 317
Livre de 7ème année de base	65	202	20 221
Livre de 8ème année de base	73	256	25 886
Total	*279*	*1 249*	*82 678*

Figure 1 : Corpus utilisé pour la conception des règles de segmentation

L'étude de ce corpus (segmenté manuellement par des linguistes) nous a permis de concevoir 183 règles de segmentation. Ces règles ont le format suivant :

Soit un marqueur déclencheur X SI le contexte gauche de X est G ET/OU SI le contexte droit de X est D ALORS prendre la décision Y (fin ou non fin d'un segment)

Figure 2 : Format de règles conçues pour la segmentation de textes (Mourad, 2001)

Ces règles peuvent être classées en trois classes relatives aux trois types de marqueurs déclencheurs à savoir les signes de ponctuation, les particules et les mots connecteurs (Baccour, Mourad, Belguith Hadrich, 2003). Nous présentons dans ce qui suit un exemple d'une règle relative à la virgule.

Contexte gauche			Marqueur	Contexte droit		
	Verbe	Espace	,			وفي صباح

SI	la virgule est suivie par un espace
ET Si	l'espace est suivi d'un verbe
ET SI	le contexte droit de la virgule commence par " وفي صباح"
ALORS	la virgule ne marque pas la fin de la phrase

C'est le cas par exemple de l'énoncé suivant :

وفي صباح مشرق من أصباح الصّيف مرّ بابن عمّه إسماعيل.

Et à une des matinées ensoleillées de l'été, il a passé à son cousin Ismail.

4 Présentation du système STAr

STAr (voir figure 3) est un segmenteur de textes arabes basé sur l'approche de segmentation

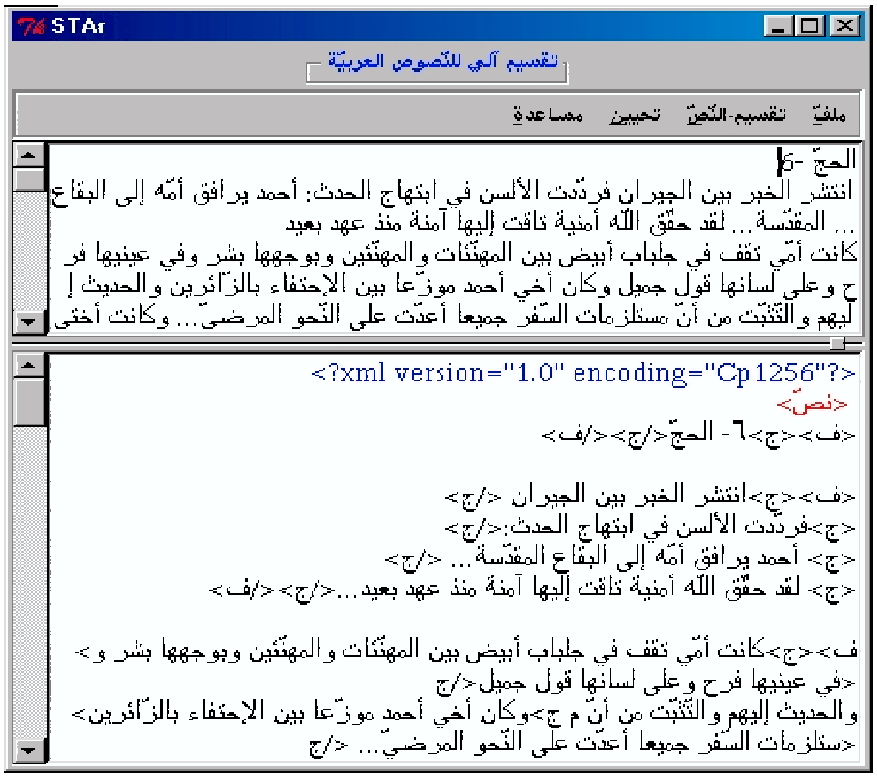

Figure 3 : Un exemple d'exécution de STAr

proposée. Il est réalisé avec le langage de programmation Perl. Il accepte en entrée un texte arabe en format txt et génère en sortie un texte segmenté en paragraphes et en phrases.

La figure 3 montre un texte segmenté par STAr. Dans le premier éditeur, figure le texte source (texte à segmenter de type .txt) et dans le deuxième éditeur figure le texte segmenté par STAr. Ce texte est généré dans un fichier XML. Les balises <نصّ> et </نصّ> indiquent le début et la fin d'un texte, les balises <ف> et </ف> représentent le début et la fin d'un paragraphe et les balises <ج> et </ج> représentent le début et la fin d'une phrase.

5 Evaluation de STAr

L'évaluation du système STAr a été réalisée sur deux corpus différents (voir figure 4).

Corpus	*Nombre de textes*	*Nombre de paragraphes*	*Nombre de mots*
Deux livres (4ème année primaire et 9ème année de base)	*144*	*991*	*40 3431*
Articles de journaux	60	510	38 062

Figure 4 : Corpus d'évaluation de STAr

Les mesures de rappel et de précision obtenues pour le premier corpus sont meilleures que ceux trouvés pour le deuxième corpus (voir figure 5). Ceci s'explique par le fait que les articles de journaux contiennent des erreurs typographiques (i.e. insertion d'un espace après la conjonction de coordination "و" (wa), omission de la lettre "الشّدّة" ('chadda), des constructions erronées, etc.) qui augmentent le taux d'erreur au niveau de la segmentation en mots, de l'identification de la catégorie grammaticale des mots et par conséquent le taux d'erreur au niveau de la segmentation en phrases augmente.

Corpus	*Rappel*	*Précision*
Livres	88.26%	80.65%
Articles de journaux	75.81%	65.66%

Figure 5 : Les mesures de rappel et de précision obtenues pour les deux corpus d'évaluation

7 Conclusion et perspectives

Dans ce papier nous avons proposé une approche de segmentation de textes arabes non voyellés qui se base sur l'analyse contextuelle des signes de ponctuation, de certaines particules et certains mots connecteurs.

Nous avons aussi présenté notre système STAr, un Segmenteur de Textes Arabes, basé sur l'approche proposée. STAr est actuellement intégré dans le système MASPAR (Multi Agent System for Parsing Arabic) d'analyse de textes arabes non voyellés (Aloulou, Belguith, Ben Hamadou, 2000), (Aloulou, Belguith, Hadj Kacem, Ben Hamadou, 2004). Ce système est composé de 5 agents (segmentation, morphologie, syntaxe, ellipse, anaphore) (Aloulou, Belguith, Hadj Kacem, Hammami, 2003). Ainsi STAr est intégré dans MASPAR en tant qu'agent pour la segmentation de textes en phrases et pourrait collaborer avec l'agent morphologie qui a pour objectif de déterminer pour chaque mot sa catégorie grammaticale ainsi que ses caractéristiques morphologiques (genre, nombre, temps, personne, etc.) (Belguith Hadrich, Ben Hamadou, 2004).

Comme perspectives, nous envisageons d'étudier la collaboration de STAr avec l'agent syntaxe. En effet, certaines ambiguïtés de segmentation ne peuvent être levées qu'à l'aide d'informations syntaxiques. De plus certaines particules utilisées dans la segmentation

peuvent à leur tour être ambiguës. Ainsi, nous envisageons de faire une étude approfondie de ces cas d'ambiguïtés.

Références

Aloulou C., Belguith Hadrich L., Hadj Kacem A., Ben Hamadou A., (2004), Conception et développement du système MASPAR d'analyse de l'Arabe selon une approche agent, *14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, du 28 au 30 janvier 2004 à Toulouse - France.

Aloulou C., Belguith Hadrich L., Ben Hamadou A., (2000), Vers un système d'analyse syntaxique robuste pour l'Arabe: Application au recouvrement des erreurs de la reconnaissance, *7ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2000)*, 16 – 18 octobre 2000, Lausanne, SUISSE.

Blachère R., Gaudefroy-Demombynes M. (1975), *Grammaire de l'arabe classique*, Éditions Maisonneuve & Larose 15, rue Victor-cousin 75005 Paris.

Beesley K. (1996), Arabic Finite-State Morphological Analysis and Generation; *COLING96*, Vol. 1, pages 89-94.

Belguith Hadrich L., Ben Hamadou A. (2004), Traitement des erreurs d'accord : une analyse syntagmatique pour la vérification et une analyse multicritère pour la correction, *Revue d'Intelligence Artificielle (RSTI- RIA)*, Hermès-Lavoisier, Vol. 18, N 5 et 6.

Belguith Hadrich L. (1999), *Traitement des erreurs d'accord de l'Arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritères pour la correction*, Thèse de doctorat en informatique, Faculté des Sciences de Tunis.

Baccour L., Mourad G., Belguith Hadrich L. (2003), Segmentation de textes arabes en phrases basée sur les signes de ponctuation et les mots connecteurs, *troisième journées scientifiques des jeunes chercheurs en génie électrique et informatique*, du 25-27 mars, Mahdia, Tunisie.

Chaâben N., Belguith Hadrich L (2003), L'étiquetage morpho-syntaxique: Comment lever l'ambiguïté dans les textes arabes non voyellés *?, troisième journées scientifiques des jeunes chercheurs en génie électrique et informatique*, du 25-27 mars, Mahdia, Tunisie.

Descles J.-P., (1997), *Systèmes d'exploration contextuelle. Co-texte et calcul du sens.* , éd. Claude Guimier, Presses Universitaires de Caen, pp. 215-232.

Debili F., Achour H., Souissi E. (2002), La langue arabe et l'ordinateur, de l'étiquetage grammatical à la voyellation automatique, *Correspondances n° 71 juillet-août 2002.*

Hammami S., Aloulou C., Belguith Hadrich L., Hadj Kacem A. (2003), Implémentation du système MASPAR selon une approche multi-agent*, IWPT'03 (International Workshop on Parsing Technologies),* 23-25 avril 2003, Nancy, France.

Mourad G. (2001), *Analyse informatique de signes typographiques pour la segmentation de textes et l'extraction automatique des citations*, Thèse de doctorat en informatique linguistique, université de Paris - Sorbonne.

Palmer D., Hearst M. (1994), Adaptive sentence boundary disambiguation, *Report No. UCB/CSD 94/797*, Computer Science Division (EECS), University of California, Berkeley, California 94720.

Silberztein M. (1993), Dictionnaires électroniques et analyse automatique de textes, Le système INTEX, Paris, Masson.