Annotation Sémantique des Énonciations en Arabe

Motasem Alrahabi, Brahim Djioua, Jean-Pierre Desclés

LaLICC – UMR 8139
Université de Paris-Sorbonne – CNRS
28, rue Serpente, 75006 Paris, France
[prenom.nom]@paris4.sorbonne.fr

Résumé

Nous présentons dans cet article un système d'annotation automatique des énonciations dans les textes en arabe. Le cas traité est celui des énonciations signées, simples ou rapportées. Le cadre théorique de l'étude linguistique des énonciations est la Grammaire Applicative et Cognitive et la méthode utilisée est l'Exploration Contextuelle. Le papier commence par l'étude linguistique de la notion d'énonciation dans les textes en arabe ; ensuite les modules fondamentaux du système informatique EXCOM sont présentés avant de finir par un exemple d'application finalisée dans la recherche d'information.

We present in this article a computational tool for semantic annotation of information in Arabic. The concerned case is the signed, simple or reported enunciations. The theoretical framework of the linguistic study of enunciations is the Applicative and Cognitive Grammar, and the method used in is the Contextual Exploration. The paper begins by a linguistic analysis of the enunciation's notion in Arabic texts. Next we deal with the automatic annotation system EXCOM. Finally, we present an example of an application in Information Retrieval system..

1. Introduction

Les applications en traitement automatique du langage, tels que l'extraction d'information, la traduction automatique et la recherche d'information nécessitent de plus en plus l'utilisation de la sémantique. Dans le cadre de ce travail, nous nous intéressons à une sémantique discursive des textes et plus particulièrement aux informations énonciatives que l'auteur laisse lors la rédaction de ses écrits. Nous considérons que les énonciations produites par l'énonciateur (énonciateur principal ou locuteurs) véhiculent dans les textes des informations importantes qui, récupérées, peuvent être utilisées par d'autres applications finalisées. Ce travail linguistique utilise la méthodologie de l'Exploration Contextuelle (EC) et est implémenté dans la plateforme EXCOM [1]. Après avoir expliqué le cadre théorique du travail, nous allons détailler l'analyse linguistique des énonciations en arabe et leurs catégorisation, présenter les marqueurs de ces énonciations et les difficultés de leurs reconnaissance automatique, et enfin nous proposerons un exemple de traitement appliqué au corpus.

2. Exploration Contextuelle

Initialisée par J.-P. Desclés [2, 3], l'EC se base principalement sur le repérage des unités linguistiques de surface qui sont des marqueurs linguistiques indépendants d'un domaine particulier. Ceux-ci sont les traces directes de l'intention énonciative de l'auteur du texte et les instruments qu'il utilise pour guider le lecteur dans son processus cognitif de compréhension. D'autres indices complémentaires peuvent agir dans le contexte afin de confirmer ou d'infirmer la pertinence du repérage. Le processus de l'EC est mis en œuvre dans un système de règles déclaratives, appelée règles d'EC, et ne fait recours à aucune analyse linguistique préalable (ex. analyse morpho-syntaxique). Afin d'illustrer le mécanisme de l'EC, prenons un exemple (*corpus Al-Jazeera*) :

ونقلت الصحيفة عن المتحدث باسم مجلس الثورة الأعلى أن أي صراع جديد في المنطقة ليس من مصلحة الولايات المتحدة, وأن "القوات الأمريكية المتواجدة في المنطقة في غاية الضعف"

et le journal a rapporté d'après le porte parole du conseil suprême de la Révolution que n'importe quel conflit dans la région n'est pas dans l'intérêt des États-unis, et que "les forces américaines qui se trouvent dans la région sont extrêmement vulnérables".

Le marqueur نقلـــت / a rapporté, est un indicateur déclencheur d'une règle d'EC. Les marqueurs عن /d'après et أن / que sont les indices complémentaires de l'indicateur qui valident les conditions de la règle.

3. Cadre théorique

La linguistique de l'énonciation introduit la notion de l'énonciateur dans l'analyse et distingue le *modus* (la manière avec laquelle le propos est présenté) du *dictum* (le contenu du propos). Une énonciation simple sera représentée par le schéma d'axiome énonciatif suivant :

DIS (modus (dictum)) JE

où le JE renvoie au sujet énonciateur et DIS à un opérateur verbal d'énonciation (ex. *je dis qu'il fait beau*). J.-P. Desclés et Z. Guentcheva [4] distinguent le *sujet énonciateur*, qui prend en charge la totalité de l'énonciation, du locuteur, « *le dernier énonciateur qui prend en charge directement la relation prédicative.* ». Ceci permet de complexifier l'expression de l'acte énonciatif :

[DIS (DIT (dictum) X)] JE & [X REP JE]

où JE renvoie au sujet énonciateur et X au locuteur (ex. *je pense qu'il doute que le voyage soit annulé*). La relation de repérage REP entre JE et X permet de spécifier dans les énoncés la relation entre les occurrences des personnes *je*, *tu* et *il* (relation qui peut-être l'identification, la différenciation ou la ruption).

4. Types d'énonciation à annoter

Dans une perspective de TAL, nous avons défini le type de l'énonciation que nous voulons repérer dans les textes de manière à ce que celle-ci soit :

- Énonciation signée : c'est une énonciation accompagnée de la trace lexicale du sujet énonciateur ou du locuteur : أجد أن الجو لطيف / *je trouve qu'il fait beau ;*
- Énonciation liée à l'activité langagière ou cognitive : lire, écrire, parler, penser, critquer, entendre, etc. ;
- Énonciation simple : أعتقد أن الوقت قد انتهى / *je crois que le temps est terminé*, ou bien énonciation rapportée (discours rapporté direct ou indirect) : أكد الوزير تمسكه بالقانون/ *le ministre a affirmé son attachement à la loi*).

5. Marqueurs linguistiques introducteurs des énonciations en arabe

L'étude linguistique menée [5] sur le corpus nous a permis de dégager les indicateurs principaux introduisant les énonciations en arabe :

- Les verbes qui introduisent nécessairement une parole, de manière directe ou indirecte (قال/ *dire*, أكد/ *affirmer*, أعلن/ *déclarer*) ;
- Les verbes qui renvoient vers ou décrivent des énonciations du même type, mais qui ne peuvent pas introduire directement une parole : سخر/*se moquer*, شجع/ *encourager*, لخص/*résumer*, etc. (...شجعني قائلاً/*il m'a encouragé en disant…*) ;
- Les participes présents dérivés des verbes du premier type, précédé d'un verbe différent (كان يمشي قائلاً/*il marchait en disant*) ;
- Les noms dérivés de verbes du premier type (نصيحة/*conseil* ; القائل/*celui qui dit*) ;
- Les syntagmes prépositionnels comme بحسب/*selon* (بحسب فلان/*selon quelqu'un*) ;

Nous nous limitons, dans cet article, à présenter le vaste cas des verbes. L'observation de ceux-ci nous a permis [6] d'introduire une classification générale sur un critère homogène : le repérage de la direction de l'information véhiculée par l'énonciation, toujours par rapport à l'énonciateur principal. Ceci a donné les cinq classes suivantes que nous considérons comme une carte sémantique de l'énonciation :

- *verbes d'émission* : dire, prétendre, diffuser, écrire, critiquer, répondre, nier, résumer, encourager, envoyer ;
- *verbes de réception* : écouter, lire, entendre, apprendre ;
- *verbes d'échange* : discuter, dialoguer, négocier, débattre ;
- *verbes transmission* : communiquer, rapporter, faire circuler ;
- *verbes réflexifs* : comprendre, douter, espérer, déduire, sentir ;

Ces classes de verbes peuvent être décrites sous forme de schèmes sémantico-cognitives [7], en considérant l'information comme une entité notionnelle qui change d'état ou qui se déplace entre les protagonistes de l'énonciation.

6. Génération du lexique

Les listes de verbes indicateurs contiennent à la base juste les *masdar*, le verbe à la troisième personne du singulier à l'accompli. À partir de ces entrées, nous voulons reconnaître dans les textes certaines formes conjuguées, non vocalisées, simples ou agglutinées. Une autre exigence est de lier à chaque verbe conjugué, la préposition liée avec, à son tour aussi, toutes les formes agglutinées compatibles (ex. أن/ أنه/ أنهم/ أنها/ أننا/ وأن/ وأنه .الخ).

Pour réponde à ces besoins, nous avons mis en place un outil basique de conjugaison de verbes. Ce générateur est basé sur la comparaison de la position et de la nature de voyelles entre les deux formes dévocalisées de l'accomplie et inaccomplie du verbe. Il s'appuie aussi sur des règles d'adaptations morphologiques et graphiques, liées à la nature du système morphologique arabe [8], comme par exemple les règles d'adaptation

entre la racine et les affixes de conjugaison, ou entre la forme conjuguée et les particules d'agglutination, proclitiques et enclitiques [9]. La sortie est au format XML, où chaque forme générée est liée aux différents types de particules rattachées. Cette opération se fait une seule fois et ne fait pas partie du processus de l'annotation.

7. Ambiguïté

Pour la reconnaissance des marqueurs dans les textes, nous sommes confrontés à l'ambiguïté provoquée surtout par la vocalisation partielle, l'agglutination et l'ordre relativement libre des mots dans la phrase [10]. D'autres ambiguïtés connexes viennent s'ajouter comme l'homographie (شرح = *disséquer /expliquer / explication*) ; la polysémie (علق= *accrocher/commenter*) ; l'anaphore et le rattachement des propositions aux verbes (تكلمت رغماً عنه الذي حدث/عن/*j'ai parlé malgré lui de ce qui s'est passé, lui et de en arabe sont homographes*).

Nous avons commencé par la dévocalisation complète du corpus, ce qui nous a permis d'avoir une approche indépendante de l'état de vocalisation des textes traités.

Ensuite nous avons mené une analyse linguistique détaillée des verbes introducteurs d'énonciation (environs 360 verbes), nous avons alors spécifié pour chacun de ces indicateurs, quels sont les indices complémentaires de désambiguïsation (en cas de besoin) et quels sont les indices qui introduisent le propos. Ceci nous a mené à formuler les remarques suivantes :

- le verbe indicateur, selon sa nature, est rattaché ou non d'une particule (préposition) : أكد أن / *affirmer que*, انتقد/*critiquer ;*
- cette particule n'introduit pas toujours le propos : على/ *sur* (نمّ على فلان قائلا… / *il a calomnié sur quelqu'un en disant…*) ;
- cette particule est parfois suffisante pour la désambiguïsation du verbe mais pas toujours (ex. ورد في/*il a été rapporté dans, ici le verbe* ورد *est ambigu malgré* في).

Les indices de désambiguïsation des verbes peuvent être de plusieurs genres :

- classe *document-thème* : مقالة / *article*, حديث/ *discussion,* برنامج/*émission* (ces indices accompagnent certains verbes ambigus en arabe comme بث/ *diffuser*, اختصر /*synthétiser*, etc.) ;
- classe *particules* : ces indices comportent notamment des propositions et d'autres *tokens* qui nous servent à lever l'ambiguïté dans certains cas (ex. دون dans دون ذكر indique qu'il ne s'agit pas du verbe ذكر/*rappeler* ou *souligner*) ;

Quant au propos, il est introduit soit directement et sans indices (يقول الحقيقة/*il dit la vérité*) soit par l'un des indices introducteurs du propos :

- classe de particules rattachées au verbe : قال أن/*dire que*, تكلم عن/*parler de ;*
- participes présents (PP) : ...سخر منه قائلاً/ *il s'est moqué de lui en disant*… ;
- noms : … أصر على رأيه بالقول/*il a insisté sur son opinion par dire*…* ;
- signes typographiques : guillemets, deux points.
- marqueurs d'ouverture : التالي, الآتي/ *le suivant, etc.*

8. Règles d'Exploration Contextuelle

Dans notre tâche, les espaces de recherche dans lesquels les règles d'EC opèrent sont les phrases. Nous avons ainsi découpé les textes en paragraphes et en phrases terminées par des points. Le principe de la segmentation automatique utilisée consiste à ne pas prendre en compte les points des non fins de phrases [11].

Les règles d'EC utilisent plusieurs moyens combinés pour repérer l'information recherchée et l'annoter, comme par exemple la présence ou l'absence de certains indices dans le contexte ou l'analyse de la position de ceux-ci dans la phrase. Elles utilisent aussi des heuristiques, adoptées après observation des résultats. Nous donnons ici à titre représentatif une règle simple d'EC :

Si un indicateur verbe se trouve dans l'espace de recherche initial
Si l'indicateur n'est pas précédé directement par une négation ou interrogation
Si l'indicateur est désambiguïsé avec les indices *document, thème, préposition*.
Si (recherche dans l'ordre)
la particule rattachée au verbe ;
OU un PP ou un nom dérivés d'un verbe indicateur ;
OU deux-points ;
se trouve dans le Contexte-Droit de l'indicateur
Alors annoter l'espace de recherche initial en tant que passage contenant une énonciation

Appliquée au corpus, cette règle doit repérer des phrases ayant la même structure typique montrée ci-dessus. Nous allons voir plus loin un extrait de l'exécution de cette règle avec le moteur EXCOM. Nous signalons ici qu'une règle d'EC peut faire appel à d'autres règles, afin d'effectuer d'autres tâches, comme par exemple dans notre cas le repérage du sujet énonciateur ou du locuteur.

9. Architecture d'EXCOM

L'implémentation informatique utilisée pour l'annotation automatique des informations énonciatives prend appuie sur la plateforme EXCOM qui s'inspire de

l'architecture modulaire GATE [12] et est décrite dans la figure suivante.

Les textes traités par EXCOM sont d'abord prétraités pour les préparer à une segmentation en phrases, paragraphes et sections. Les conditions de déclenchement des règles d'annotation sont exprimées de différentes façons qui déclenchent certains niveaux du moteur d'annotation. Chaque niveau fait appel à un algorithme général de fonctionnement.

Ce moteur est construit sous une forme multicouche où chaque brique répond à un besoin d'annotation particulier. Le module REGEX fait appel à un moteur d'expressions régulières. Avec le support d'Unicode, l'extraction d'information peut se réaliser sur des documents multilingues. Le module d'exploration contextuelle (EC) est composé :

- d'un ensemble de marqueurs linguistiques (indicateurs et indices) ;
- d'un ensemble de règles d'EC qui se présentent sous la forme de règles déclaratives.
- d'un moteur d'EC qui applique les règles en respectant la primauté de l'indicateur sur les indices complémentaires.

Le résultat de l'application de ces règles est un texte annoté. Les annotations sont des marques sous forme d'éléments et attributs XML. La sémantique de ces annotations est liée à l'organisation de la catégorie du point de vue reconnue par le système EXCOM. L'objectif de cette plateforme est de proposer une exploration du texte afin de l'augmenter d'informations sémantiques sous forme d'annotations sémantiques discursives.

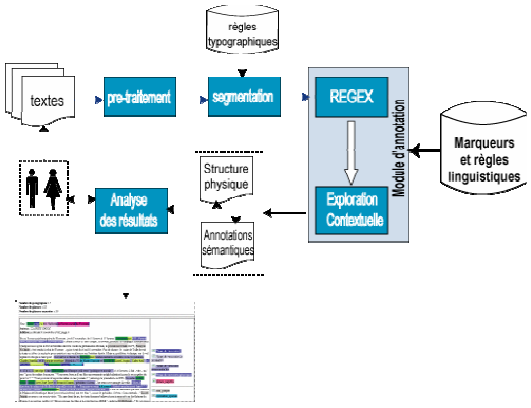

Figure 1 : Architecture informatique d' EXCOM

Exemple d'annotation : Voici une sortie de texte annoté (Figure 2), sur lequel nous avons appliqué l'exemple de règle décrite ci-dessus. Le texte est tiré de notre corpus qui contient environ mille articles de presse en langue arabe. Les sources sont : Al-Jazeera, Al-Nahar, Al-Alam, AL-Ahram, Al-Sabah.

ردود متباينة على برنامج مراقبة مسلمي أميركا
أثار كشف مجلة "يو.أس نيوز أند وورلد ريبورت" عن برنامج اتحادي واسع لمراقبة عشرات المساجد والمواقع التي يرتادها المسلمون في الولايات المتحدة، ردود أفعال متباينة. **وأوضحت** المجلة في موقعها الإلكتروني **أن** أكثر من 120 موقعا يتردد عليها مسلمون من مساجد وأماكن سكن ومتاجر ومستودعات وضعت تحت المراقبة، وذلك بحثا عما تعتبره الإدارة الأميركية قنابل نووية محتملة. هذا وأن عمليات المراقبة التي يقوم بها مكتب التحقيقات الفدرالي (FBI) وفريق متخصص في الطاقة النووية كانت تشمل في بعض الأحيان قيام عناصر حكومية بالدخول إلى ممتلكات بدون مذكرات تفتيش أو أمر من المحكمة، في عمل يرى محامون أنه غير قانوني. كما أن هذا البرنامج الحكومي الذي يشمل العاصمة واشنطن ومدن شيكاغو وسياتل وديترويت ونيويورك ولاس فيغاس، مصنف على أنه سري للغاية وقد بدأ بعد أحداث 11 سبتمبر/أيلول 2001.
وحسب الصحيفة فإن بعض المراقبين العاملين في إطار البرنامج تعرضوا للتهديد بالطرد من عملهم في حال سؤالهم عن قانونية البرنامج. ويدخل البرنامج ضمن خطة عامة وافق عليها البيت الأبيض تقضي بتخويل وكالة الأمن القومي مهمة المراقبة الإلكترونية للأهداف الأميركية المحتملة دون الحاجة إلى أوامر قضائية.
ردود متباينة
وقد تراوحت ردود المسؤولين الأميركيين بين الصمت المطبق نظرا لحساسية الموضوع وبين التأكيد على شرعية عمليات المراقبة بصفة عامة لاستباق أعمال تستهدف الأمن القومي. وقد نفت وكالة الأمن الوطني تركيز العمليات على أماكن وممتلكات خاصة بعينها تعود لأفراد.
ومن جهتها انتقدت الجالية المسلمة هذا البرنامج لتغذي الانتقادات الموجهة للبيت الأبيض حول قرار التنصت على المكالمات الهاتفية. وقد **عبر** مجلس العلاقات الأميركية-الإسلامية **عن** مخاوفه **من أن** يجعل برنامج التجسس على المسلمين الأميركيين من هذه الفئة "كبش فداء". **وقال** المدير التنفيذي للمجلس نهاد عوض "أخوف من أن نكون ننتقل إلى دولة خوف تجعل الأقليات مثل المجموعات المسلمة الأميركية كبش فداء".
سجال التنصت
ويأتي كشف هذا البرنامج وسط السجال الذي تعيشه الولايات المتحدة حول مشروعية التنصت الذي أذن فيه الرئيس الأميركي جورج بوش على الاتصالات الهاتفية والإلكترونية لمواطنين أميركيين بدون إذن قضائي. ومن الجدير بالذكر أن جورج بوش ما لبث أن **دافع عن** هذا المخطط **قائلاً** "العمليات التي أمرنا بها من دون إذن رسمي من القضاء بعد أحداث الحادي عشر من سبتمبر/ أيلول 2001 لم تكن تستهدف إلا أولئك الذين كانت لهم علاقات معروفة مع القاعدة"
كما يتزامن ذلك مع النقاش الحاد الذي يدور بشأن مستقبل قانون مكافحة الإرهاب المعروف بقانون الوطنية (باتريوت آكت) الذي يتيح للرئيس "استخدام كل القوة اللازمة والمناسبة" لمكافحة من يسمون الإرهابيين.

Figure 2 : extrait d'un texte annoté

10. Perspectives

Nos objectifs à moyens termes se focalisent sur les modalités de l'énonciation et sur l'étude des autres catégories d'énonciation et leur représentation avec des schèmes sémantico-cognitifs. En même temps, nous sommes en train d'effectuer de nombreux tests sur la plate forme EXCOM afin de valider les règles d'EC et d'effectuer une évaluation du système. L'application finalisée de ce travail sera une sorte de stratégie de recherche d'informations qui fait appel à l'organisation discursive des textes. Les éléments de l'annotation discursive (phrases annotées, énonciateurs et contenus de l'énonciation), telles qu'elles sont présentés dans cet

article, doivent subir un processus d'indexation (avec Lucene www.apache.org/lucene). L'application de recherche d'information sémantique consiste donc à répondre à la question « Qui Dit Quoi ? », L'utilisateur pourra choisir le profil de l'énonciateur : énonciateur principal, locuteur, ou bien locuteur précis (saisie d'une entité nommée). Il peut aussi choisir une sous catégorisation de la carte sémantique (émission, réception, échange…). Par défaut, le système va rechercher toutes les énonciations de tous les énonciateurs.

11. Bibliographie

[1] Djioua B., Flores J. G., Blais A., Desclés J.-P., Guibert G., Jackiewicz A., Le Priol F., Leila N.-B., Sauzay B., « EXCOM : an automatic annotation engine for semantic information », *FLAIRS 2006*, Floride

[2] Desclés J.-P., Jouis C., Oh H-G., Reppert D., « Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte ». In *Knowledge modeling andexpertise transfer*, pp.371-400, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam, 1991.

[3] Desclés J.-P., *Systèmes d'Exploration Contextuelle. Co-texte et calcul du sens*, (ed. Claude Guimier), Presses Universitaires de Caen, p. 215-232, 1997.

[4] Desclés J.-P., Guentcheva Z., « Enonciateur, Locuteur, Médiateur dans l'activité dialogique », in *Colloque International des Américanistes*, Quito, Equateur, 1997.

[5] Alrahabi M., Mourad G., Djioua B., « Filtrage sémantique de textes en arabe », *JEP-TALN 2004*, Fès.

[6] Alrahabi M., A. H. Ibrahim, J.-P. Desclés, « Semantic Annotation of Reported Information in Arabic » , *FLAIRS 2006*, Floride.

[7] Desclés J.-P., *Langages applicatifs, langues naturelles et cognition*, Hermès, Paris, 1990.

[8] Dichy J., « On lemmatization in Arabic, A formal definition of the Arabic entries of multilingual lexical databases ». In *Proceedings of the Workshop on Arabic Language Processing : Status and Prospects*, Toulouse, 2001.

[9] Aloulou C., Belguith L. H., Kacem A. H., Ben Hamadou A., « Conception et développement du système MASPAR d'analyse de l'arabe selon une approche agent », *RFIA 2004*, Toulouse.

[10] El-Kassas D. et Kahan S., « Modélisation de l'ordre des mots en arabe standard », *JEP-TALN 2004*, Fès.

[11] Mourad G., « La segmentation de textes par Exploration Contextuelle automatique, présentation du module SegATex », *Inscription Spatiale du Langage : structure et processus ISLsp*., Toulouse, 2002.

[12] Cunningham H., Maynard D., Bontcheva K., Tablan V. « GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications ». *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02).Philadelphia, July 2002