Une approche linguistique pour l'extraction des connaissances dans un texte arabe

Houda Saadane
LIDILEM, Université Stendhal – Grenoble III, 1180, avenue central, F-38400 Saint Martin d'Hères
houda.saadane@e.u-grenoble3.fr

RÉSUMÉ

Nous présentons dans cet article un système d'extraction de connaissances en arabe, fondé sur une analyse morphosyntaxique profonde. Ce système reconnaît les mots simples, les expressions idiomatiques, les mots composés et les entités nommées. L'analyse identifie aussi les relations syntaxiques de dépendance et traite les formes passives et actives. L'extraction des connaissances est propre à l'application et utilise des règles d'extraction sémantiques qui s'appuient sur le résultat de l'analyse morphosyntaxique. A ce niveau, le type de certaines entités nommées peut être révisé. L'extraction se base, dans nos expérimentations, sur une ontologie dans le domaine de la sécurité. Le RDF (Resource Description Framework) produit est ensuite traité pour regrouper les informations qui concernent un même événement ou une même entité nommée. Les informations ainsi extraites peuvent alors aider à appréhender les informations contenues dans un ensemble de textes, alimenter une base de connaissances, ou bien servir à des outils de veille.

ABSTRACT

A linguistic approach for knowledge extraction from an Arabic text

We present in this paper a knowledge extraction system for Arabic. The information extraction is based on a deep morphosyntactic analysis. It also recognizes single words, idiomatic expressions, compounds and named entities. The analysis also identifies dependency relations, verb tenses and passive/active forms. Information extraction is application-independent and uses extraction rules that rely on the result of the morphosyntactic analysis. At this level, some named entity categories can be reconsidered. This extraction is based in our experimentations on the security ontology. The Resource Description Framework (RDF) obtained is then processed to gather information concerning a single event or named entity. The information extracted can help to understand the information contained in a set of texts, to infer knowledge into a knowledge base, or be used for monitoring tools.

MOTS-CLÉS : Analyse linguistique, fouille de textes, arabe, entités nommées, extraction d'informations, règles d'extraction, ontologie.

KEYWORDS : Linguistic analysis, Text Mining, Arabic, named entities, information extraction, extraction rules, ontology.

1 Introduction

Les évolutions rapides des nouvelles technologies sont accompagnées d'un essor important

de la quantité d'information disponible sur le Web, et nécessitent le développement d'outils pour analyser et structurer les documents textuels. Ainsi, les documents en arabe sur le Web, à l'instar des auteurs langues, se multiplient en nombre, en contenus et en quantité. Pour traiter cette grande masse de textes, une possibilité est de recourir à des outils de fouille de textes ou d'extraction de connaissances. Pourtant, dans ce domaine en pleine émergence, les outils qui permettent d'analyser les documents arabes se limitent en général à l'extraction d'entités nommées.

L'objectif de cet article est de présenter un système d'extraction des connaissances pour l'arabe qui, après avoir effectué une analyse linguistique profonde du texte, extrait les entités nommées, et les relie à des événements, en se basant sur une ontologie métier. Le développement d'un tel système exige l'étude des propriétés des données textuelles en soulevant essentiellement les problèmes concernant l'analyse et la représentation des contenus des textes (Cherfi., 2002).

Dans cet article, nous commençons par présenter dans la section 2 un état de l'art du domaine de l'extraction de connaissances, suivi dans la section 3 d'une description des étapes de l'analyse linguistique profonde. La section 4 décrit le processus d'extraction sémantique qui se base sur les résultats de l'analyse syntaxique et sur des concepts et des règles d'extraction spécifiques au domaine traité. Dans la section 5, nous présentons le traitement qui consiste à aligner les informations recueillies au niveau du document pour regrouper les occurrences qui désignent une même entité nommée, ou un même événement. Nous exposons dans la section 6 les premiers résultats que nous avons obtenus. Nous terminons par une ouverture sur les perspectives d'amélioration de notre système.

2 Etat de l'art

De nombreux travaux de recherche ont abordé la question de l'extraction de connaissances, citons TEMIS (TExt MIning Solutions) qui est un éditeur de logiciels d'enrichissement sémantique des contenus. Sa plateforme Luxid (Kuznik et al., 2010) propose des fonctionnalités d'extraction d'information qui sont réalisées par le moteur Insight Discoverer Extractor. Ce serveur d'extraction d'information pour l'analyse des textes enchaine trois étapes : l'analyse de corpus (identification de la langue), l'analyse linguistique réalisée par Xelda et l'extraction des connaissances (identification des entités nommées, reconnaissance des relations entre les entités). Cette étape repose sur la technologie Skill Cartridge (Brun et al., 2009) qui propose un ensemble de règles et de composants linguistiques définissant l'information à extraire. La cartouche va permettre de réaliser une analyse sémantique pour fournir les relations sémantiques. Ces relations sémantiques font référence par la suite à des entités nommées et à des patrons spécifiques au domaine de l'intelligence économique. Un rôle est assigné à chaque entité, afin de définir sa position dans la relation sémantique. Les langues analysées par IDE, sont le français, l'anglais, l'espagnol, l'italien et l'allemand.

Le projet SAMAR[1] (Station d'Analyse Multimédia en langue Arabe) est, quant à lui, destiné aux journalistes travaillant en langue arabe. Son objectif est le développement d'une plateforme de traitement multimédia à destination de la presse et des médias arabes. Parmi les principales composantes de ce projet, il y a l'analyse sémantique et l'extraction des

[1] http://www.samar.fr/

entités nommées, qui constituent les sujets des informations à analyser. Ces entités sont par la suite stockées dans une base de connaissances.

Concernant plus particulièrement l'extraction des connaissances en langue arabe, les travaux se sont focalisés sur la reconnaissance et l'extraction des entités nommées. Parmi ces travaux, nous pouvons mentionner les travaux de (Zitouni et al., 2005) qui utilisent des techniques d'apprentissage automatique (Modèles de Markov à Entropie Maximale) en considérant des jeux de descripteurs idoines. L'utilisation d'un corpus parallèle pour l'extraction des entités nommées en arabe a été adoptée par (Samy et al., 2005). Cette méthode est aussi basée sur des règles, mais avec en plus l'utilisation d'un lexique monolingue de langue espagnole pour permettre l'extraction des entités nommées en espagnol. Une fois que ces entités sont extraites, un processus de transcription en arabe est appliqué sur ces entités.

Enfin, le travail de thèse de Mesfar (2008) avait pour objectif une analyse morpho-syntaxique et une extraction des entités nommées en arabe standard. Son système est basé sur une combinaison des résultats obtenus par le biais d'un analyseur morphologique et de grammaires locales représentant des règles d'identification écrites à la main. D'autres travaux récents se basent sur des méthodes à base de règles Shaalan et al., 2009 ; Zaghouani et al., 2010.

Les études que nous avons décrites ci-dessus proposent une solution à la problématique de l'extraction des entités nommées, mais n'abordent pas, ou peu, le sujet de l'extraction de connaissances relatives à ces entités. Notre objectif est de proposer un système d'extraction de connaissances pour l'arabe, qui sera capable de repérer les entités nommées, mais aussi les relations sémantiques qui les relient, pour un domaine particulier modélisé dans une ontologie. Notre analyse est fondée sur la technologie des automates d'états finis.

3 Analyse linguistique profonde

L'analyse linguistique profonde est nécessaire pour assurer une extraction d'informations sûre, pertinente et complète, par exemple en reliant des éléments qui peuvent être éloignés dans la phrase initiale.
L'analyse que nous avons mise au point se divise en plusieurs étapes allant du découpage en mots jusqu'aux relations que ceux-ci entretiennent au sein d'une phrase. Les principales étapes de cette analyse sont décrites dans les sous-sections suivantes :

3.1 Découpage en mots

La tokenisation permet le découpage du texte en mots, les « tokens », séparés par des ponctuations ou par des espaces. Elle prend aussi en compte les balises, les dates abrégées, etc. Citons l'exemple de la tokenisation en mots de la phrase باريس مدينة الجن والملائكة . » (Paris la ville des diables et des anges) donnera : باريس | مدينة | الجن | و | الملائكة. C'est une étape qui va permettre d'attribuer ensuite à chaque token des catégories et des propriétés sur lesquelles portera l'analyse profonde.

3.2 Analyse morphologique

Le travail de l'analyseur morphologique consiste à retrouver la forme de surface d'un mot

stocké dans le lexique à partir de la forme canonique de ce dernier (infinitif du verbe, masculin singulier d'un adjectif, etc...). Cette étape est primordiale lors de l'analyse linguistique. Elle se divise à son tour en plusieurs étapes : la consultation du dictionnaire des formes fléchies d'une part pour récupérer la normalisation du mot et d'autre part, pour permettre de récupérer les informations linguistiques (genre, nombre, catégorie grammaticale, etc.) de ce mot. L'une des particularités de la langue arabe est la présence des formes agglutinées (formes avec des proclitiques et des enclitiques). Ces formes ne sont pas présentes dans le dictionnaire des formes fléchies. Pour identifier ces formes et les traiter correctement, nous avons ajouté un segmenteur de clitiques (proclitiques et enclitiques) à l'analyse morphologique. Cette segmentation des formes agglutinées se déroule de la manière suivante (Semmar et al., 2005) :

1. Recherche de toutes les compositions possibles entre les clitiques (proclitique, enclitique) et le radical en utilisant les dictionnaires des proclitiques, enclitiques et formes fléchies.

2. Chaque radical est ensuite recherché dans le dictionnaire des formes fléchies. Si ce radical n'existe pas dans le dictionnaire, des transformations morphologiques sont appliquées avant leur suffixation en se basant sur des règles de réécriture, enfin le radical résultat est de nouveau recherché dans le dictionnaire des formes fléchies. Par exemple, considérons la forme agglutinée «بسيارته» (avec sa voiture) et les clitiques inclus dans cette forme (ب, ه). Le radical récupéré «سيارت» n'existe pas dans le dictionnaire des formes fléchies. Mais après l'application de la règle de réécriture transformant la lettre «ت» en «ة» en fin de mot, le radical modifié «سيارة» (voiture) est trouvé dans le dictionnaire des formes fléchies et la forme agglutinée «بسيارته» est découpée en proclitique + radical + enclitique comme suit : بسيارته = ب + سيارة + ه (avec sa voiture).

3. Une étape supplémentaire permet de vérifier la relation d'ordre au sein d'une représentation des formants du mot sur un vecteur ordonné (Zmantar et al., 2009). La principale propriété de celui-ci est que chaque proclitique est incompatible avec un proclitique de même position, en raison de la relation d'ordre strict qui régit les formants du mot graphique. Exemples : wa et fa coordonnants (واو العطف et فاء), qui occupent tous les deux la même position sur le vecteur d'ordre, sont incompatibles entre eux (ils ne peuvent pas apparaître dans un même mot). Cette étape doit aussi vérifier les règles, syntaxiques mais aussi sémantiques, de compatibilité et d'incompatibilité entre les proclitiques et les enclitiques.

Cette analyse reconnaît aussi des expressions idiomatiques afin de grouper certains mots pour les considérer comme une seule unité (سكة الحديد : Chemin de fer). Cette reconnaissance se fait à l'aide de règles et de dictionnaires.

Si, après ces étapes, un mot reste inconnu, le système lui attribue une (des) catégorie(s) par défaut, en s'appuyant sur des informations révélées par sa forme de surface. Par exemple, s'il s'agit d'un mot en caractères latins majuscules, comme ONU, il sera étiqueté comme un nom propre.

Après cette analyse morphologique, et particulièrement pour le traitement de la langue arabe, la majorité des mots restent ambigus à cause de l'absence des voyelles courtes arabes

dans les textes (Debili et al., 1998), ce qui est moins prononcé pour les autres langues. Le problème majeur rencontré dans cette phase est celui de l'ambiguïté lexicale et grammaticale, qui découle du fait que lorsqu'un mot est reconnu, l'analyseur morphologique peut fournir plusieurs interprétations qui renvoient à plusieurs catégories syntaxiques ou à plusieurs sens. Le rôle du désambiguïseur morpho-syntaxique qui intervient par la suite, est de réduire le nombre des ambiguïtés grammaticales en utilisant des matrices de désambiguïsation. Ce sont des matrices de bi-grammes et tri-grammes de catégories obtenues à partir d'un corpus étiqueté du LDC (Arabic Treebank), et désambiguïsé manuellement. Le résultat de l'application des n-grammes nous permet d'obtenir la suite de couples mot-catégories la plus probable. L'ambiguïté lexicale est conservée à ce niveau, pour être traitée plus tard, au niveau de l'extraction sémantique.

3.3 Repérage des dates

Lors de l'analyse morphologique, un traitement spécifique intervient pour le repérage des dates. Ceci permet ensuite à la désambiguïsation d'être plus efficace, étant donné que les dates ne sont plus constituées d'une suite de catégories, mais sont associées à une catégorie « date ».

Les dates et heures se composent de l'indication normalisée du temps qu'elles représentent. Nous nous sommes basés sur la norme ISO 8601 avec le format AAAAMMJJ où AAAA représente l'année sur 4 chiffres, MM représente le numéro du mois, sur 2 chiffres et JJ représente le quantième dans le mois, sur 2 chiffres. Par exemple, «أفريل 1984 (Avril) 01» est normalisé de la manière suivante : «19840401 », «أفريل (Avril) 1984 » est normalisé par «198404XX», «غداً : demain» est normalisé par «XXXXXX+1».

3.4 Mots composés

Une étude linguistique spécifique de la langue arabe nous a permis de définir et d'écrire un certain nombre de règles dans le but d'établir des relations de dépendance (contiguës et non contiguës) entre les mots au sein du syntagme nominal. Ces relations permettent ensuite de reconnaître les mots composés présents dans une phrase.

Citons l'exemple de «أرملة الشهيد» (la veuve de martyr), nous avons une relation entre deux mots associés par annexion (معرف بالإضافة), qui relie le mot indéfini)veuve(أرملة et le mot défini الشهيد (martyr) pour donner une relation de type "NomRelNom".

3.5 Relations sujet-verbe-complément

Nous avons défini un certain nombre de règles, issues d'une étude expérimentale pour l'identification et le repérage des relations syntaxiques dans une phrase. Notons que certains verbes demandent un complément, contrairement à d'autres. Ces verbes sont appelés des verbes transitifs. Il faut définir la liste des verbes transitifs et des verbes intransitifs, étant donné que, en arabe, la position des mots ne suffit pas à en déduire la fonction syntaxique du mot. Les voyelles courtes l'indiquent, mais elles ne sont généralement pas indiquées dans les textes écrits. C'est pour cela qu'il faut se baser sur la transitivité ou sur la non transitivité du verbe pour déterminer quelles sont les relations qui existent entre un nom et un verbe.

Voici les relations que nous détectons :

- Les relations agent-verbe, qui permettent d'identifier l'agent de l'action (pour répondre à la question : qui a fait l'action?)
- Les relations verbe-complément, qui permettent d'identifier qui a subi l'action, ou encore les circonstanciels qui nous renseignent sur le moyen (comment? Avec quoi?), la date (quand?), le lieu (où ?), ... de l'action.

3.6 Passif

Afin de rendre l'étape de construction des règles d'extraction des connaissances plus efficace, l'analyse linguistique profonde adopte en interne la même représentation pour une phrase passive, et pour son équivalent à la forme active. Cette phase consiste donc à identifier les formes passives et à les transformer en formes actives. Voici quelques structures syntaxiques exprimant un passif (Ziad, 2010) :

- Passif avec un verbe doublement transitif. مُنِح الشاعرُ جائزةً : On a accordé un prix à l'écrivain
- Passif avec un verbe transitif indirect, précédé par une préposition, حُكِمَ عليه بالاعدام : Il a été condamné à mort.
- Emploi de tournures modernes du passif, qui expriment le complément d'agent : مِنْ قِبَل ، مِنْ طَرَفِ ، مِنْ جَانِب ، عَلَى يَد (par).

Dans l'exemple suivant : إعتقلت الفتاة على يد الشرطة (la fille a été arrêtée par la police), pour ne pas confondre entre la personne qui fait l'action et la personne qui la subit, il est important de savoir si la forme est active ou passive. Ici, la forme est active mais emploie une tournure moderne du passif qui exprime le complément d'agent (عَلَى يَد), donc le sujet est la police et le complément est la fille. Nous obtenons donc les relations suivantes :

- relation agent-verbe entre إعْتَقَلَ (arrêter) et شُرْطَةٌ (Police) reliés par le mot عَلَى يَد (par)
- relation verbe-complément entre إعْتَقَلَ (arrêter) et فَتَاة (fille).

```
<relation reltype="SV">
 <head>
  <posBeg>112</posBeg>
  <lemma>إعْتَقَلَ </lemma>
  <catPos index="no">+verbe</catPos>
  <prop
index="no">+vbpassif+acc+3fs</prop>
  <posEnd>118</posEnd>
 </head>
 <dept>
  <posBeg>133</posBeg>
  <lemma>شُرْطَةٌ</lemma>
  <catPos index="no">+nom</catPos>
  <prop index="no">+fs</prop>
  <posEnd>139</posEnd>
 </dept>
 <lingIndication index="no">
  <posBeg>126</posBeg>
```

```
<relation reltype="VC">
 <head>
  <posBeg>112</posBeg>
  <lemma>إعْتَقَلَ </lemma>
  <catPos index="no">+verbe</catPos>
  <prop
index="no">+vbpassif+acc+3fs</prop>
  <posEnd>118</posEnd>
 </head>
 <dept>
  <posBeg>119</posBeg>
  <lemma>فَتَاة</lemma>
  <catPos
index="no">+annppers</catPos>
  <prop index="no">+pers+fs</prop>
  <posEnd>125</posEnd>
 </dept>
</relation>
```

```
  <lemma>عَلَى يَد</lemma>
  <catPos index="no">+prepN</catPos>
  <prop index="no">+passif</prop>
  <posEnd>132</posEnd>
 </lingIndication>
</relation>
```

TABLE 1 – Exemple d'extraction des relations syntaxiques dans une phrase passive.

Head : unité qui constitue la tête de la relation
Dept : unité qui constitue le dépendant de la relation
LingIndication : balise qui contient des indications sur les unités qui permettent de relier les termes d'une relation, et qui serviront lors de l'extraction sémantique.

3.7 Reconnaissance des entités nommées

Cette phase consiste à mettre en œuvre un système de reconnaissance et de typage des entités nommées. Dans notre approche, nous avons opté pour un système à base de règles linguistiques qui exploitent l'étiquetage syntaxique, des marqueurs lexicaux (déclencheurs) et des dictionnaires de noms propres. La mise en place de règles de reconnaissance d'entités nommées a nécessité une recherche profonde sur certains traits linguistiques propres aux entités nommées en arabe.

Exemple : " الأخ مُعز غرسلاوي " (le frère Moez Garsallaoui) Dans cet exemple, nous avons le titre de civilité "الأخ : frère" suivi d'un prénom et d'un nom propre. Voici la représentation que nous obtenons :

```
<en entype="pers">
 <relation reltype="AnnpNP">
  <head>
   <posBeg>1104</posBeg>
   <lemma>غرسلاوي </lemma>
   <catPos index="no">+np</catPos>
   <posEnd>1111</posEnd>
  </head>
  <dept>
   <posBeg>1092</posBeg>
   <lemma>أَخٌ</lemma>
   <catPos ndex="no">+annppers</catPos>
   <prop index="no">+pers+ms</prop>
   <posEnd>1097</posEnd>
  </dept>
 </relation>
```

```
<relation reltype="PrenomNP">
  <head>
   <posBeg>1104</posBeg>
   <lemma>غرسلاوي </lemma>
   <catPos index="no">+np</catPos>
   <posEnd>1111</posEnd>
  </head>
  <dept>
   <posBeg>1098</posBeg>
   <lemma>مُعِز</lemma>
   <catPos
index="no">+prenom</catPos>
   <prop index="no">+m</prop>
   <posEnd>1103</posEnd>
  </dept>
 </relation>
</en>
```

TABLE 2 – Exemple de reconnaissance des entités nommées de type Personne.

4 Extraction sémantique

L'entrée de cette étape est constituée de la sortie de l'analyse morpho-syntaxique décrite précédemment. Cette analyse fournit les informations suivantes :

- les lemmes des mots ainsi que leur position dans le texte, et leur catégorie grammaticale
- les relations de dépendance syntaxique entre les mots,
- les entités nommées typées.

Nous avons choisi une représentation interlingue en anglais de toutes les informations, dans le but de faciliter la lecture des informations extraites par les non arabophones et de faciliter la fusion d'informations provenant de documents en plusieurs langues. Nous avons eu recours à deux types d'opérations : l'utilisation de dictionnaires de traduction existants et l'ajout d'un système de translittération pour les entités nommées qui n'existent pas dans les dictionnaires (Saadane et al., 2012).

L'extraction de connaissances permet de mettre en évidence des entités nommées et des relations relatives à un concept particulier, par exemple : « arrestation », « attentat », « condamnation », « construction ». Le déroulement de cette étape s'effectue en trois temps : la sélection des concepts potentiellement présents, la sélection des règles à appliquer, puis l'application des règles.

4.1 Sélection des concepts probables

Une étape primordiale lors de l'extraction des connaissances consiste à repérer les déclencheurs. Ces déclencheurs peuvent être des mots, des expressions ou des relations, et indiquent qu'une relation relative à un concept peut être présente dans le texte. Les déclencheurs sont présentés sous forme de deux colonnes :

- la première colonne contient les mots, les expressions ou encore les relations qui indiquent la présence du concept dans le texte traité.
- la deuxième colonne définit le concept (arrestation, transfert, émission, union,...) associé à l'élément de la première colonne.

إعْتَقَلَ Arrest
VC#نَقَلَ#رِسَالَة# Emission

Comme nous l'avons mentionné et comme le montre l'exemple précédent, les déclencheurs peuvent être des mots (إعتقل, زواج) :(mariage, interpeller…), des expressions ou bien des relations (VC# نَقَلَ # رِسَالَة #) : (VC#transmettre#message#).

Il est nécessaire qu'un déclencheur soit présent dans l'entrée afin d'être en mesure d'extraire l'information présente. A partir d'un déclencheur, un concept est obtenu ce qui nous permet ensuite de sélectionner les règles à appliquer.

4.2 Sélection des règles à appliquer

Les déclencheurs nous ont permis d'obtenir la liste des concepts présents dans le texte. Ces concepts nous amènent alors vers une liste de règles qui vont être confrontées aux relations issues de l'analyse syntaxique. Si les relations syntaxiques correspondent aux règles définies, elles pourront alors être extraites. La définition des règles à appliquer comporte aussi deux

colonnes :

- La première colonne indique les concepts. Ils sont ensuite comparés aux concepts probables sélectionnés à la phase précédente par le biais des déclencheurs
- La deuxième colonne liste les règles spécifiques à un concept pouvant être appliquées. Ces règles contiennent des relations syntaxiques (SV, VC) entre un verbe et une liste de catégories, avec d'éventuelles prépositions.

Arrest SV#إعْتَقَل#<en># عَلَى يَد (Arrest SV#arrêter#<en># par)
Arrest SV#إعْتَقَل#<en># (Arrest SV#arrêter#<en>#)
Arrest VC#إعْتَقَل|#<en># (Arrest VC#arrêter#<en>#)

Pour illustrer notre propos, prenons l'exemple suivant : إعتقلت العروض على يد الشرطة (Elaroud a été arrêtée par la police). Lors de la première étape, le mot «اعتقل» (arrêter) a été repéré comme déclencheur du concept « Arrest» (arrestation). Ce concept est associé aux règles présentées ci-dessus. Afin de pouvoir être sélectionnées, les règles doivent correspondre à une relation syntaxique présente dans la phrase. Or, dans la phrase « Elaroud a été arrêtée par la police », « arrêter » a un complément « Elaroud » de type entité nommée, et « arrêter » a un sujet « police », étant donné que la phrase est au passif. Donc, ce sont les règles SV#إعْتَقَل #<en>#عَلَى يَد et VC#إعْتَقَل #<en># qui seront sélectionnées.
A ce stade intervient un traitement qui permet l'application des règles sélectionnées lors de la phase précédente aux relations syntaxiques effectivement présentes, afin d'extraire l'information en question.

4.2.1 L'application des règles sélectionnées

Les règles sélectionnées à l'étape précédente sont appliquées aux relations syntaxiques afin d'en extraire les connaissances présentes. Les connaissances sont extraites à partir de la sortie de l'analyse linguistique profonde. La règle d'extraction indique ensuite quels sont les éléments qui doivent être extraits, et quelle est leur sémantique. Si nous reprenons l'exemple « إعتقلت العروض على يد الشرطة: Elaroud a été arrêtée par la police», l'une des règles sélectionnée est la relation verbe-complément entre «arrêter» et une entité nommée. Cette règle indique que le concept extrait sera «Arrest» (Arrestation), dont le patient est l'objet de «arrêter», c'est-à-dire «Elaroud». Le résultat obtenu sera alors de la forme suivante : *<gs:Arrest rdf:nodeID="id17Arrest"><wn:undergoer rdf:nodeID="id1Elaroud"/></gs:Arrest>*

4.2.2 L'extraction des entités nommées et son contrôle

Toutes les entités nommées détectées au niveau de l'analyse linguistique sont extraites en conservant leur type : « Personne », « Lieu », « Organisme », « Mesure », « Date », « Produit » ou encore « Inconnu ». Mais l'analyse linguistique a pu se tromper sur le type d'une entité nommée. Cette étape effectue un contrôle sur le type des entités nommées extraites.

Le contrôle consiste à vérifier l'adéquation entre le type de l'entité nommée issu de l'analyse linguistique, et le type de l'entité nommée proposé par la règle. Les types des entités nommées peuvent être modifiés si la règle d'extraction considère que le type de l'entité nommée est incompatible avec le type de l'entité nommée issu de l'analyse linguistique.

L'exemple suivant illustre ce phénomène d'incompatibilité : « الجزأير : ُان باريس ُاعلنت *Paris déclare que l'Algérie* ...». Au niveau morpho-syntaxique, Paris est considéré comme une entité nommée de type «lieu», mais dans le cadre de l'action d'émission d'un message, l'agent

ne peut pas être un lieu. En fait, l'émission ne peut être réalisée que par une personne ou une organisation et si à l'origine elle a été considérée comme la capitale de la France, la nouvelle catégorie est une organisation, et nous pouvons en déduire que c'est le gouvernement français.

4.2.3 La création d'entités

Il peut arriver qu'une règle fasse référence à une entité sans que celle-ci existe. C'est le cas lorsque l'entité nommée n'a pas pu être repérée au niveau de l'analyse syntaxique, ou bien lorsqu'il ne s'agit pas d'une entité nommée comme dans l'exemple «أدين : il a été condamné». Pour faire apparaître le patient de l'action, la règle d'extraction va créer l'entité manquante :

*VC#*أَدَانَ *#***<$pers>**# → *<en entype="pers">***<$pers>**</en>

Notons que l'extraction et/ou la création d'entités peuvent introduire des ambiguïtés. Une relation peut demander comme objet ou sujet une entité de type lieu ou organisation. Ces ambiguïtés sont alors générées, pour être résolues plus tard, grâce à un contrôle manuel par exemple, ou bien grâce à la mise en cohérence. Cependant, lorsque l'ambiguïté se situe entre les types personne ou organisation, le type « agent », qui est générique, sera indiqué.

5 Mise en cohérence

Précisons pour commencer que les résultats de l'extraction de connaissances est un graphe **RDF** faisant référence aux concepts et propriétés issus de l'ontologie intégrée dans le système. Cette étape va permettre de rassembler les informations concernant une même entité nommée, ou une même action. Elle permet aussi d'exploiter les métadonnées attachées au document. La construction de notre système d'extraction a nécessité la définition des informations d'intérêt dans le domaine de sécurité.

Nous avons choisi, pour cela, de développer une ontologie du domaine qui servira de guide aux différentes étapes d'extraction. La diversité des documents exploités nécessite que l'ontologie soit assez générale tout en contenant des concepts et des propriétés spécifiques au domaine de la sécurité.

Nous avons développé une ontologie interne, en nous basant sur des ontologies existantes telles que ***foaf*** pour les agents (Person et Organization) et en ajoutant d'autres concepts décrivant les connaissances que nous souhaitons extraire tels que les actes de violences, les déplacements, les transferts d'argent… La construction de cette ontologie a été réalisée manuellement à base de corpus. A l'heure actuelle, notre ontologie compte 106 classes et 200 propriétés d'objets.

5.1 Regroupement des entités nommées

L'un des problèmes des différentes étapes d'extractions réside dans le fait que les graphes obtenus peuvent contenir des duplications inutiles de nœuds. Ce phénomène est particulièrement visible pour les entités nommées que l'on retrouve à plusieurs reprises dans un même document. L'objectif de cette étape consiste à regrouper les différentes occurrences d'une même entité nommée sous un même et unique URI. Ce problème est généralement connu sous le nom de 'Record linkage' ou 'Entity resolution' et a été abordé par différentes approches (Elmagarmid et al., 2007).

Dans le contexte d'un graphe RDF, et dans le domaine de l'extraction sémantique, nous adoptons une méthode basée sur un ensemble de règles. Ces règles ont été définies pour identifier les entités nommées dupliquées et permettre leur regroupement. Citons un exemple de ces règles : deux personnes sont identiques dans un même document, si elles ont le même nom et prénom, et qu'il n'y a pas d'autres informations contradictoires, par exemple «junior» et «senior».

5.2 Résolution des dates relatives

Parmi les problèmes que l'analyse linguistique ne résout pas il y a les dates relatives. Ces dates ne sont pas toujours exprimées d'une manière explicite dans les textes. Pour résoudre ce phénomène, nous nous appuyons sur les trois aspects suivants :

1. La représentation adoptée par l'ontologie : l'ontologie décrit chaque date comme un intervalle. Elle contient donc les attributs suivant : (1) **dtstart :** date de début, (2) **dtend :** date de fin, (3) **type :** le type de calendrier utilisé, qui correspond à des constantes prédéfinies dans l'ontologie (grégorien, arabe, chinois …), (4) **authorValidation :** donne une indication sur quand a eu lieu l'action, si cette dernière se situe dans le passé ou le futur, grâce notamment aux temps des verbes liés à la date, (5) **day :** le jour de la semaine, lorsqu'il est précisé.
2. la sortie de l'analyse linguistique
3. les métadonnées du document analysé (notamment la date d'édition du document).

La sortie de l'analyse linguistique nous permet d'identifier les occurrences où la date extraite est incertaine. Dans le contexte de la presse écrite, il est fréquent d'extraire des dates relatives à un jour de la semaine ou à une indication dans le temps. Par exemple un événement devant se dérouler «نهاية الأسبوع المقبل» : le week-end prochain » pour un article paru le lundi 01 avril 2013 (une métadonnée du document). Les métadonnées sont alors exploitées pour définir une date incertaine se situant entre le samedi 06/4/2013 et le dimanche 07/4/2013.

6 Premiers résultats

Pour estimer l'efficacité de notre système, nous avons mené deux types d'évaluation : une évaluation quantitative concernant la phase de segmentation et la phase d'extraction d'entités nommées et une évaluation qualitative (intrinsèque) de l'extraction de connaissances.

Nous avons comparé nos performances de segmentation avec l'outil de Stanford[2] en apportant quelques modifications aux résultats pour pouvoir les comparer avec ceux de notre outil. Par exemple, dans l'outil de Stanford, l'article défini fait partie du mot, contrairement à notre segmenteur qui considère que l'article défini est un token indépendant.

Nous avons calculé la précision sur un ensemble de documents (articles de presse Aljazeera) segmentés par l'outil de Stanford et corrigés manuellement. Nous avons eu une précision de 0,98% avec notre segmenteur contre 0,96% avec l'outil de Stanford.

[2] http://nlp.stanford.edu/projects/arabic.shtml

Pour évaluer notre approche d'extraction d'entités nommées nous avons réalisé nos expériences sur le corpus ANER[3] (Benajiba et al., 2007) qui est composé de 150 000 occurrences de mots. Ce corpus distingue les types d'entité nommée suivantes : lieu (**LOC**ation qui représente 30.4% des EN observées), personne (**PERS**on : 39%), organisation (**ORG**anization : 20.6%) et une classe qui regroupe toutes les autres EN, de type « divers » (**MISC**ellaneous : 10%). Nous nous sommes intéressés à la reconnaissance des trois premiers types des entités nommées et avons obtenu une précision de 89,05% pour la détection des entités nommées de type personne, 91% pour les lieux et 83.41% pour les organisations.

Nos systèmes de segmentation et d'extraction des entités nommées obtiennent de bons résultats. Par ailleurs, notre système présente encore quelques faiblesses comme le montre la précision pour les entités de types «organisation».

L'absence ou la non disponibilité des outils et des travaux de référence dans le domaine de l'extraction des connaissances pour le traitement de l'arabe a été un vrai obstacle pour mesurer la performance de notre système, et ne nous permet pas de comparer notre approche avec les autres travaux. C'est la raison pour laquelle nous avons lancé des phases de tests afin d'améliorer et de compléter l'extraction d'informations.

Une question engendrée par cette phase est : sur quel corpus peut-on tester notre module ? Nous avons opté pour les corpus suivants :

- Corpus de textes sur Malika El-Aroud. Ce corpus est assez général et regroupe une grande partie des concepts présents dans notre ontologie.
- Ensemble de corpus propres à chaque concept, composés d'articles journalistiques. Ces corpus ne sont pas généraux mais peuvent permettre d'étudier et d'améliorer en profondeur un type de concept. Les corpus spécifiques sont les suivants : « arrestation », « attentat », « condamnation », « construction », « décès », « divorce », « émission », « mariage », « paiement », « rencontre » et « transfert ».

Afin de recouvrir un maximum de cas tout en améliorant la reconnaissance et l'extraction d'information, l'utilisation conjointe de ces deux types de corpus paraît être la meilleure solution.

Citons l'exemple suivant : وفي مرحلة لاحقة إقترنت أُم عُبيدة بالأخ مُعز غرسلاوي من أصول تونسية وانتقلت معهُ لسويسرا. (À un stade ultérieur Umm Obeyda s'est mariée avec le frère Moez Garsallaoui d'origine tunisienne et elle s'est installée (partie) avec lui en Suisse.)

Pour cet exemple, notre système extrait les connaissances suivantes:

- entités nommées de type Personne : Umm Obeyda et frère Moez Garsallaoui.
- entités nommée de type Lieu : Suisse
- concept «Union» extrait grâce au verbe «إقترن» «se marier», avec deux bénéficiaires : «Umm Obeyda » et « Moez Garsalloui».
- concept « Transfert » est extrait grâce au verbe «انتقل».

Voici la représentation des connaissances extraites, dans notre outil de visualisation :

[3] http://users.dsic.upv.es/~ybenajiba/downloads.html

FIGURE 1 – Représentation des connaissances extraites, dans notre outil de visualisation.

Conclusion

Nous avons décrit dans cet article un système d'extraction des connaissances dans des textes arabes, basé d'une part sur une analyse linguistique profonde, et d'autre part sur une extraction sémantique utilisant une ontologie du domaine. L'évaluation effectuée sur ces premiers travaux nous a permis de déceler globalement la qualité de notre extraction mais aussi de donner naissances à d'autres problématiques à étudier.

Notre approche à base de règles contextuelles atteint l'état de l'art pour l'extraction d'entités nommées en arabe. Notre méthode d'extraction de connaissance montre le caractère indispensable d'une analyse syntaxique profonde dans le repérage de telles informations.

Pour rendre notre système d'extraction plus complet, nous allons étendre l'analyse syntaxique, en y ajoutant la recherche des antécédents des anaphores présentes dans les textes. En effet, si les pronoms, utilisés fréquemment dans les textes pour éviter les répétitions, ne sont pas liés à l'entité à laquelle ils font référence, nous risquons de perdre beaucoup des informations présentes dans les textes. Il faut noter que les limites des systèmes linguistiques et statistiques actuels nous orientent vers une future combinaison de ces approches pour une meilleure extraction. Nos travaux futurs s'orientent vers une extension de notre système à d'autres domaines.

Remerciements

Je tiens à remercier l'Agence Nationale de la Recherche portant la référence ANR-09-CSOSG-08-01, pour son aide qu'elle nous a apportée pour mener à bien ce travail, ainsi que Mme Aurélie Pradelles-Rossi et M Christian Fluhr.

Références

BENAJIBA, Y., et ROSSO, P. (2007). ANERsys 2.0 : Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. *In Proc. Workshop on Natural Language-Independent Engineering, 3rd Indian Int. Conf. on Artificial Intelligence, IICAI-2007,* Pune, India, December 17-19.

BRUN, C., DESSAIGNE, N., EHRMANN, M., GAILLARD, B., GUILLEMIN-LANNE, S., JACQUET, G., KAPLAN, A., KUCHARSKI, M., MIGEOTTE, A., NAKAMURA, T. et VOYATZI, S. (2007). Une expérience de fusion pour l'annotation d'entités nommées. *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis.

CHERFI, H. (2004). Etude et réalisation d'un système d'extraction de connaissances à partir de textes. *Thèse de doctorat*, novembre 2004, LORIA, Nancy.

DEBILI, F. et ACHOUR, H. (1998). Voyellation automatique de l'arabe. *In Proceedings of the Workshop on Computational Approaches to Semitic Languagues.* Montreal. Canada, pages 42–49.

DEBILI, F. et SOUISSI, E. (1998). Étiquetage grammatical de l'arabe voyellé ou non. *Correspondance de l'IRMC, N°71*. Tunis.

ELMAGARMID, A.K., Ipeirotis, P.G. et VERYKIOS, V.S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and data Engineering (TKDE).* 19(1) pages 1–16.

KUZNIK, L., GUÉNET, A-L., PERADOTTO, A., et CLAVEL, C. (2010). L'apport des concepts métiers pour la classification des questions ouvertes d'enquête. *Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montréal, Canada .

MESFAR, S. (2008). Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. *Thèse de doctorat*, novembre 2008.

MIKATI, Z. (2010). Du Data Mining au Sense mining : modèle pour une analyse de la langue arabe, et ses représentations formelles en vue d'une application à des données demandant une haute sécurité. *Thèse de doctorat se*, mai 2010.

SÂADANE, H., ROSSI, A., FLUHR, C. et GUIDÈRE, M. (2012). Transcription of Arabic Names into Latin. *Actes the 6th international conference SETIT 2012 (Sciences of Electronic, technologies of Information and Telecommunications).* March 2012. Sousse, Tunisia.

SAMY, D., MORENO, A. et MA GUIRAO, J. (2005). A proposal for an Arabic named entity tagger leveraging a parrallel corpus. *In Proceedings of International Conference on Recent Advances on Natural Language Processing* RANLP '05. Borovets.

SEMMAR, N., GARA, F. et FLUHR, C. (2005). Linguistic resources and analysis for unvowelled Arabic text processing in information retrieval. In *Actes de 2nd International Conference on Machine Intelligence, ACIDCA-ICMI-2005, Tozur (Tunisia)*, 5-7 Novembre 2005.

SHAALAN, K. et RAZA, H. (2009). NERA : Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology,* 60(9) : pages 1652–1663.

ZAGHOUANI, W., POULIQUEN, B., EBRAHIM, M. et STEINBERGER, R. (2010). Adapting a resource-light highly multilingual named entity recognition system to arabic. *Proceedings of the Seventh conference on International Language ressources and Evaluation (LREC'10),* pages 563–567.

ZITOUNI, I., SORENSEN, J., LUO, X. et FLORIAN, R. (2005). The impact of morphological stemming on Arabic mention detection and coreference resolution. *In Proceedings of Workshop on Computational Approaches to Semitic Languages,* pages 63–70, Ann Arbor, Michigan.

ZMANTAR, Y. et DICHY, J. (2009). L'analyse automatique des mots-outils en arabe.*2ème conférence Internationale – Systèmes d'information et Intelligence Economique 2009.* Hammamet,Tunisia.