Semantic Annotation of Reported Information in Arabic

Motsem Alrahabi, Amr Helmy Ibrahim, Jean-Pierre Desclés
LaLICC (Langage, Logique, Informatique, Cognition et Communication)
UMR 8139, Université Paris – Sorbonne, CNRS
28, Rue Serpente 75006 Paris – France
Tél. : (33) 01 53 10 58 25
{motasem.alrahabi, amr.ibrahim, jean-pierre.descles}@paris4.sorbonne.fr

Abstract
In this article, we are presenting a semantic annotation tool for Arabic texts with a strategy adapted to the automatic location of reported information. The method used is that of Contextual Exploration, which consists of using purely linguistic knowledge to identify semantic-discursive textual representations.

This work has been carried out in the framework of a thesis at the LaLICC Laboratory at the University Paris IV- Sorbonne.

Keywords: Semantic annotation, Arabic language, Contextual Exploration, linguistic markers, ambiguity.

Introduction
The automatic identification of Reported Information (RI) is a particularly important problem in the domain of automatic summarization, technological or economic watches etc. This consists of carrying out a semantic annotation of the information searched in a text, by specifying the author of this information, his position in relation to his remarks, and where or when he said them.

However, the automatic location of RI, like other semantic-discursive mechanisms in a text, is a task that is even more delicate because the concerned linguistic procedures implicated in its formation differs according to the language. This idiosyncrasy is particularly clear in Arabic where work on the automatic treatment of texts regularly comes up against difficulties related to the morphology of the language and to its syntactic system [Aloulou et al. 03, Gaubert 01, Jaccarini 97, etc.]. We especially mention the phenomenon of agglutination [Debili 01], the absence of vocalisation or the mixed order of words in the sentence [el-Kassas 04].

The method of Contextual Exploration (CE) offers the advantage of getting past certain difficulties of traditional automatic treatment of texts, for example the difficulties met with in morpho-syntactic analysis. Initialized by J.-P. Desclés [Declés 91, 97], CE process is set up in a system of declarative rules and it is principally based on a surface analysis of the context by finding linguistic indicators independent of a particular domain. These markers are the direct traces of the enunciative intention of the author of the text and the instruments he uses to guide the reader in his cognitive process of comprehension. Then we need to confirm or to declare null the pertinence of the location using complementary indicators that are present or not in the context. This approach has been the occasion for different computer applications, such as automatic summarization [Berri 96, Minel 02], extraction of causal relationships [Jackiewicz 98] and relationships between concepts [Cartier 04], segmentation [Mourad 01] etc.

The following is an example of automatic annotation of a conclusive remark (Al-Jazeera Corpus):

Finally, the doctor concluded his speech by saying: I think according to my diagnostics that the president’s operation will take place in the next few days.

This sentence is annotated by a CE rule thanks to the principal indicator (اختتم / concluded) and the secondary clue (قال / by saying).

We will now explain the different steps for constructing the necessary linguistic resources relevant to the automatic identification of reported information. Next, we will evoke the implementation of this data. We will finish with a presentation of tests and perspectives for the future.

1. Management of Linguistic Resources
The methodology used in the construction of a Context Exploration (CE) system follows a sequence of ordered steps.

1.1 Specification of the framework: Reported Information
Reported Information is any part of the text which allows the enunciator to report on the words or actions of someone else. Reported speech is thus part of this notion. It is therefore appropriate to distinguish between several levels: that of the primary enunciator, that of the speaker and, according to the verb, that of the interlocutor. We can of course have nested reported information.
The enunciator or the speaker can more or less take responsibility for his speech [Desclés et al. 97]. The responsibility taken indicates the degree of the enunciator’s commitment regarding his remarks. For example, the verb to affirm does not have the same meaning as the verb to doubt though both are RI indicators. In this article, we only treat certain aspects of this act of language that is Reported Information (RI); the research about the remaining problems is still going on. As a result, we do not treat the principal enunciator or the nested reported information; we do not treat the taking of responsibility of the enunciator or the speaker regarding his remarks. We will thus explain today, in an operational perspective of NLP, how to identify the speaker, «the last enunciator who takes directly responsibility for the predicative relationship. » [Desclés et al. 97] and his remarks in the text: Who Says What? Consequently we will show the phases of conception of this point of view.

1.2 Definition of the Corpus

Our experiences are based on a corpus of Arabic language texts made up of more than a thousand press articles from two sources:

• The complete corpus from the newspaper Le Monde Diplomatique (LMD) in Arabic.
• A corpus of articles and press releases from the Arabic channel Al-Jazeera.

1.3 Linguistic Analysis of the Corpus

To start with, this part allows us to unravel the textual representations linked to reported information and to isolate the relevant linguistic markers [Alrahabi et al. 04]. These markers are the triggers of the CE process. The RI can be introduced by linguistic markers, typographical markers [Mourad et al. 01] or a combination of both. Linguistic markers are often verbs of communication such as to announce, nouns such as response or adverbs such as according to. We have about 280 verbs and about a hundred nouns and locutions acting as triggers of the RI task. There is indeed a quite tight correlation between the morpho-syntactic property and the semantic property of nearly four hundred verbs involved in reported speech as well as in RI, as has been showed for French by Maurice Gross [Gross 1975] and for Arabic by A. H. Ibrahim [Ibrahim 1979]. But since the methodology of CE doesn’t rely, for many reasons that will not be discussed here, on morpho-syntactic analysis, we will not explore this path here.

Secondly, we need to resolve the discursive ambiguity of the trigger indicator by exploring its context in search of other complementary clues. These can be linguistic, as in the case of named entities or coordinating particles. The clues can also be typographical (quotation marks, colons, footnotes towards bibliographical references) or positional (place of the complementary that in the context that follows the verb). We point out at this stage the impact of the presence or absence of a complement like that (Inna class) with the verb. In reality, it becomes rather difficult (especially in Arabic) to locate the speaking subject and his remarks in the absence of this clue (mainly when the quotation marks and the colons are also absent). Example (Al-Jazeera Corpus):

عسكري على طول ساحل قطاع غزة...

The Israeli authorities declared on Friday night 08/03/1996 the set up of a military circle all along the coast of the Gaza Strip...

1.4 Organization and Mode of Access of Linguistic Markers

The indicators and the linguistic clues are organized in separate groups according to grammatical and enunciative criteria. Thus, the lexical category makes the first important distinction: the verbs are separated from the nouns or the particles. Next, concerning the verbs, we started classifying them in empirical and provisional groups. Other more detailed criteria for classification will be applied later to all the verbs, especially concerning the modality and the responsibility that the enunciators take for their remarks.

1.4.1 Verb classes: Here is a non-exhaustive list of some classes of RI verb indicators:

- **Declaration class**: /أعلن, صرح/ to declare, /أقول, أذكر, أقترح, أترجم/ to inform, etc.
- **Observation class**: /لاحظ, وجد/ to observe, /فهم, حس/ to feel, etc.
- **Explanation class**: /شرح, بحث, سبق, حوافر/ to explain, /بالنسبة, برسخ/ to prove, /على استبدال/ to analyse, etc.
- **Negotiation class**: /رسائل, تبدأت/ to discuss, /تلقى, تنفيذ, تفاوض/ to negotiate, /تناوب/ to discuss, etc.
- **Continuation class**: /أضاف, أكمل/ to continue, /تابع/ to follow up, etc.
- **Summarization class**: /أختتم, ازايد/ to sum up, /يختصر, إلى النهاية/ to reduce, etc.

The lists only contain the simple forms of the markers, meaning the masculine singular form of the accomplished verb and the masculine singular form of nouns. In order to get around a morpho-syntactic analysis of the texts, we will carry out an automatic generation of certain inflected and agglutinated forms of linguistic markers. Then we will proceed with their recognition in the texts.

1.4.2 Generation of the Glossary: We have set up an automatic glossary generator for Arabic, adapted to the semantic annotation of the reported information. Only the
part we consider the most important and which concerns
the verbs has been carried out so far. As a reminder, the
Arabic glossary is composed of roots, a root corresponds to
a notion (example: رقص/to dance). Based on a single root, all
the other concepts linked to a notion are derived
according to some schemes. A graphic word in Arabic is
formed [Cohen 70] from:

\[ \text{proclitic} + \text{prefix} + \text{base} + \text{suffix} + \text{enclitic} \]

Certain particles in Arabic are linked to simple forms to
constitute a single graphic block, where the ambiguity can
be potentially difficult to resolve.

The verb generator is based on an algorithm of detection of
morphological variations between the two forms of the
accomplished and unaccomplished of the verb.

It is necessary to specify the kind of suffix particles that
can be agglutinated to verbs. Thus we naturally distinguish
three kinds of suffixes in Arabic:

• Those that accept the agglutination of particles referring
to remarks, like to say (كلم فلان عن شيء: كلمة عن
الشيء);

• And finally those that don’t accept this kind of
agglutination (عبر عن الشيء لفلان: عبر عنه لفلان).

The generator is then able to give the complete conjugation
of the verb according to the parameters that we have
chosen. Thus each verb will be conjugated according to the
following grammatical traits:

• Gender: masculine or feminine;
• Number: singular, dual and plural;
• The third person, which corresponds to the speaker.

Agglutination concerning the particles attached to the
generated forms, the particles of coordination (ex. ف و
(ب) or (د) and, according to the entered verb, the suffixes
(ه) or (ن) (هم وهم).

Example: For the verb أكذب/to affirm we will have the
following generated forms:

أكذب، وأكذب. وأكذب، وأكذب، وأكذب، وأكذب، وأكذب، وأكذب.
ويكذب، ويُكذب، ويُكذب، ويُكذب، ويُكذب، ويُكذب، ويُكذب، ويُكذب، ويُكذب.

He affirmed, and he affirmed, so he affirmed, he
affirmed it, and he affirmed it, so he affirmed it, he
affirms, and he affirms, so he affirms, he affirms it,
and he affirms it, and he affirms it, etc.

The linguist can also provide other information at the time
of verb generation, mainly the nature of Inna type particle.
In fact, not all the verbs of a list are followed by the same
type of complementary clue, so we have found it more
reliable to directly attach each verb entry to its own
conjunction (Inna class).

Example:

1 - قال إن الزيد قد خرج
2 - قال إن الزيد قد خرج
3 - قال إن الزيد قد خرج
4 - قال إن الزيد قد خرج
5 - قال إن الزيد قد خرج
6 - قال إن الزيد قد خرج
7 - قال إن الزيد قد خرج

If we consider that the introductory verb and its Inna
always exist, we can then describe all the attested cases by
the language according to the following structure:

<table>
<thead>
<tr>
<th>Reading Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Right_Context</strong></td>
</tr>
<tr>
<td>After_Inna</td>
</tr>
<tr>
<td>(from the indicator’s point of view)</td>
</tr>
</tbody>
</table>

This offers us three search fields that the CE rules will
exploit. We must also find the other models for the cases
where, for example, the Inna particle does not exist or

1.5 Conception of the CE rules

The role of a CE rule is to offer a semantic filtering
strategy adapted to the automatic annotation. Once the
trigger indicator is located in the text, the next step is to
resolve any ambiguity concerning this indicator by
exploring its context through the search of secondary clues.
Hence the capital importance is the definition of the
indicator’s context.

1.5.1 The search field: The notion of search field is
strictly linked to the task treated. As an example, the search
field for the location of enumerations is not the same as
that for thematic announcements or defining expressions.
Concerning the identification of RI, the smallest unit we
can work with is a textual passage containing one (or more
than one) verbal or nominal predication. In practice, this
represents a textual fragment containing one or more
trigger indicators. Then the fragment is cut into several
segments each containing a single indicator verb with other
elements coexisting in the context. Taking a simple search
field, containing a single indicator verb, we can then study
all the possibilities of positioning of the elements we are
looking for in this field, according to the nature of the verb
and its Inna type particle. To do this, we must establish a
certain organization of the sentence in Arabic in the case of
reported information. So that the RI has informative value,
we exclude imperative, interrogative and negative
sentences. Consequently, in the case of an active
elementary structure we can have several syntactic
constructions such as:

1 - (he) said that spring is coming
2 - said Zaid that spring is coming
3 - (he) said today that spring is coming
4 - today (he) said that spring is coming
5 - Zaid said that spring is coming
6 - Zaid said today that spring is coming
7 - today said Zaid that spring is coming
when the RI introducer is a noun or a term. So it already allows us to specify the placement of the sought after elements. Next, we must find the sought after elements more precisely, with the help of other heuristics and a secondary group of clues like “named entities” or others. If the results of these searches are positive, the semantic annotations are attributed to the sentence in question, to the speaker, and to his remarks.

1.5.2 Declaration of the Rules: A CE rule is formalized by heuristics in a declarative form, made up of conditions and actions concerning the attribution of semantic labels. Example of a basic rule:

<table>
<thead>
<tr>
<th>CE rule # 3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Given a sentence P</td>
</tr>
<tr>
<td>If (indicator exists in P)</td>
</tr>
<tr>
<td>If (Inna exists in right_context)</td>
</tr>
<tr>
<td>If (left_context is empty)</td>
</tr>
<tr>
<td>If (before_Inna is not empty)</td>
</tr>
<tr>
<td>Then:</td>
</tr>
<tr>
<td>If (before_Inna contains a speaker of the class Named_Entities)</td>
</tr>
<tr>
<td>If (remark is in the field (after_Inna to the end of the sentence) or in the field (after_Inna to next_Indicator))</td>
</tr>
<tr>
<td>Then:</td>
</tr>
<tr>
<td>Give a semantic annotation to the sentence</td>
</tr>
<tr>
<td>Give a semantic annotation to the speaker</td>
</tr>
<tr>
<td>Give a semantic annotation to the remark</td>
</tr>
</tbody>
</table>

This rule allows us to pick up a sentence like the one in example 2.

In the premises of the CE rules we can use several means to take away the ambiguity. For example, the verification of the morphology of the marker (the linking suffixes and prefixes), the presence or absence of certain words in the context, the distribution of words in the sentence, the placement of a word in the sentence, etc. We also use lists of secondary clues, like lists of named entities (names of places, time expressions, titles and functions, etc.), lists of thematic terms, lists of certain grammar particles etc.

2. Computer Implementation of Linguistic Resources

The goal behind the conception of this tool is to have a coherent and complete environment adapted to our needs in which we can build and exploit linguistic knowledge. Three essential points have thus been respected: firstly, the separation of linguistic data from the computer implementation; secondly, the ease of use (the linguist doesn’t need to have advanced knowledge of computers); and thirdly, the use of standard formats of exchange (XML) in order to facilitate the exchange of data and the mobility of the tool. This tool offers several functions for the preprocessing and processing of the data:

2.1 Preprocessing of the Data

2.1.1 Cleaning: In order to treat a text in our system, it must first be converted into raw text format to delete any layout, and to encode it in UTF-8.

2.1.2 Devocalisation: The collection of texts is generally partially vocalized. Our choice is to eliminate all of the vocalizations that exist in the texts, meaning the ten signs [Zaghibi 02] that mark the Arabic pronunciation. This is because in most arab countries, today’s Arabic texts are only partially vocalized, especially in the newspapers. So the system can function for all kinds of texts without worrying about the problem of vocalisation. For example: the devocalisation of a word like صرح Ito declare would give صرح.

2.1.3 Segmentation: Unlike Latin languages, the segmentation of Arabic texts cannot be simply done using typographical signs [Mourad 02]. Added to this are other difficulties like the absence of capitals letters in Arabic or the ambiguity of the agglutinated conjunction to the words that follow (، / and، ＼so) [Baccour et al. 03]. The technique that we have adopted allows us to cut the text into paragraphs and sentences. Each sentence ends with a period or starts back on a new line. Several rules to clarify the period have been encoded in the program in the form of regular expressions. The output is in XML format. We feel that the results of this technique go perfectly with the principle of CE. A first “general” segmentation gives whole sentences ending with periods; and according to the annotation profile (the RI in our case) the first step of filtering will offer a finer segmentation of the sentence, based on semantic criteria that we have seen above.

2.2 Treatment of Data: Semantic Annotation

Three rounds of filtering are necessary for this strategy of contextual exploration, and only the first two have been carried out so far.

During the first round of filtering we locate the verb indicators and their Inna type clues as well as typographical markers : or “”. This allows us to construct the search field in which the CE rules will function. The presence of trigger indicators will start the second round of filtering, and rules will test their conditions in the search fields that are already outlined based on adequate heuristics and secondary groups of clues.

The premises of the CE rules are encoded in the tag of the XML files. As a rule is being read, the tags are looked through; the conditions are verified and the actions are executed. Finally, the third round of filtering is dedicated to resolving the conflicts between the results of several rules on the same passage of the text, to assigning degrees of pertinence to the locations and to choosing the modes of visualization and textual navigation.
3. Tests and partial evaluation

The tests carried out on the generation, segmentation, first and second round of filtering seem satisfactory. But we cannot provide any reliable figures at this point until the final results can be obtained and evaluated, as the other tasks to be completed aren’t yet fully studied and implemented.

3.1 Example of processing

Let’s take an example to show the functioning of the different steps of execution, with one type of rule (described below). Starting with the generator, we will first generate the necessary forms of a RI introductory verb, for example "قَالَ" / "to say:"

The list of generated forms is coded in XML files in which each child corresponds to a form of indicator. The Inna type secondary clue is added as the attribute of each indicator tag. Here is an extract from the XML generated file:

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<liste nomFichier="indicateursVerbes.xml">
  <indicateur particule="ن" annotation="RI">
    <element nom="قائمة"/>
  </indicateur>
  <indicateur particule="ن" annotation="RI">
    <element nom="فَعَلَ"/>
  </indicateur>
  <indicateur particule="ن" annotation="RI">
    <element nom="قَالَ"/>
  </indicateur>
</liste>
```

Now let’s consider for instance a corpus like that of Al-Jazeera. We can choose a text (or several texts) and we can carry out the semantic annotation. In this way we first convert the text in raw text format, then we devoicealize and segment it.

Now we will launch the filtering process with some CE rules. We search the text for indicators and Inna clues. Here is an extract showing different possible options after going through the first round. We can see that three sentences are kept and each one has several search fields: sentences 1 and 2 of the paragraph 6 and sentence 2 of the paragraph 7:

```
```

In the second round, each time an indicator is found in the text, the CE rules are triggered and executed one after another. Here is an extract of the result:

```
```

Translation:

```
```

Only two sentences will be kept by our rule. The second one (sentence number 2, paragraph number 6) will not be kept because its indicator is not surrounded with any
secondary IR clue. The two sentences that are kept will have a semantic annotation for the speaker and his remark. We have to notice that in the right context of the sentence chosen from the extract above, we resorted to a list of named entities that helped the localisation of the clue (الرئيس بوش).

3.2 Difficulties and Solutions

The Arabic language presents several difficulties for automatic treatment due mainly to the absence of a complete vocalisation, the absence of capital letters, the agglutination and the relatively free order of words in a sentence. The contextual analysis done by the CE allows us to clear up different kinds of ambiguity, and to provide a semantic annotation of the textual passages containing reported information. Nevertheless, certain difficulties demand a pertinent linguistic analysis and the writing of efficient CE rules. For example, in the case of anaphora or the precise location of the subject, a problem closely linked to the recognition of the named entities.

4. Conclusion and Perspectives

In this paper we have presented a method of semantic annotation and automatic identification of RI in Arabic texts. The method used, CE, has allowed us for the most part to get rid of the difficulties of the automatic treatment of Arabic and to attribute the semantic annotation to the involved textual passages. The CE methodology adopted offers us the advantage of avoiding self implication as it happened with the classic procedures of NLP, mainly concerning morpho-syntactic analysis or statistical methods.

We have three short-term objectives: First, to formalize the structures corresponding to the other cases of RI seen above, for example, those that do not contain an Inna clue or those introduced by a noun. Secondly, to study the responsibility taken by the enunciators for their remarks. And finally, to technically carry out the third round of filtering. This will allow us to make an evaluation. Subsequently, it is possible to integrate this tool, from a technical and conceptual point of view, in a more important setting, where other tasks can be carried out.

References


Mourad, G. 2002. La segmentation de textes par Exploration Contextuelle automatique, présentation du module SegATex ; In Inscription Spatiale du Langage : structure et processus ISLsp., Toulouse.