Arabic NooJ Parser: Nominal Sentence Case

Nadia GHEZAIEL HAMMOUDA & Kais HADDAR

Miracl Laboratory, University of Sfax, Tunisia
Outline

- Introduction
- Previous work
- Typology of nominal sentence
- Proposed method
- Experimentation and evaluation
- Conclusion and perspectives
Introduction

Parsing Arabic corpora is an important task aiming to:

• Understand Arabic language
• Enrich and enhance the electronic resources
• Increase the efficiency of application for Arabic corpora like translation or recognition of named entities
Arabic is considered one of the difficult languages to analyze due to its morphological and syntactical characteristics.

There are two types of sentences in Arabic:

- verbal sentence
- nominal sentence
Different forms of nominal sentence exist

Interaction between nominal and verbal sentences

Formalization of rules requires much effort to guarantee several qualities:

• efficiency
• robustness
• extensibility
Introduction

Transducers have proved their usefulness in a wide variety of applications in NLP.

Transducer cascades made possible to carry out a robust and a highly precise syntactic analyzes on different corpora.
... Introduction

Transformation of recursive graph of transducers into transducer cascade is very interesting.

The transformation is a difficult task due to:

- difference between application levels
- interaction of linguistic phenomena
For the cascade:

- The order of transducers should respect different constraints
- Constraints can be deduced from studies done on Arabic corpus
Our objective is to construct an Arabic parser implemented in NooJ.

To do that, we:

• study, essentially, the Arabic nominal sentences but also other sentence forms

• establish a set of rules recognizing nominal sentences that can be generalized to treat any sentence type

• implement the transducer cascade in NooJ
## Previous work

<table>
<thead>
<tr>
<th>Approach</th>
<th>Formalism</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habash et al., 2009</td>
<td>Hybrid Production rules and numeric algorithms</td>
<td>Arabic parser</td>
</tr>
<tr>
<td>Abdelkarim et al., 2006</td>
<td>Linguistic HPSG</td>
<td>HPSG parser for nominal sentences</td>
</tr>
<tr>
<td>Boukédi et al., 2014</td>
<td>Linguistic HPSG</td>
<td>HPSG Parser for Coordination</td>
</tr>
<tr>
<td>BenSalem et al., 2017</td>
<td>Hybrid Property grammar</td>
<td>Arabic Parser for ATB sentences</td>
</tr>
</tbody>
</table>
... Previous work

- Lack of complete linguistic approach for all sentence type
- Absence of NooJ Arabic parser for sentence complex structure
- Parser execution time is very high
Typology of nominal sentence

- Nominal sentence
  - Topic
  - Attribut
... Typology of nominal sentence

Form of topic

Single word

Mariam [is] beautiful

Phrase

The door of the garden’s home is beautiful

باب حديقة المنزل جميل
... Typology of nominal sentence

Forms of attribut

Single word

الباب صغير

The door is small

Phrase

الباب صغير الحجم

The door is small in size

Sentence

الباب هو المدخل الوحيد

The door is the only entrance
A nominal sentence can contain a verbal sentence with different forms:

- **Verb**
  - The girl runs
  - الفتاة تجري

- **Verb + adjective**
  - The girl running quickly
  - الفتاة تجري سريعا

- **Verb + noun**
  - The man helps poor people
  - الرجل يساعد المحتاجين

- **Verb + noun + adjective + noun**
  - The director gives prices to distinguished students
  - المديرة تمنح التلاميذ المميزين الجوائز
Proposed method

Step 1
- Corpus
- Segmentation process
- XML file
- NooJ
- 2017
- 16
- Dictionaries
- NooJ Parser

Step 2
- Constraints
- Priorities
- Dictionaries
- Syntactical representations
High level of granularity is used to solve some lexical ambiguities and reduce the complexity of parser.

**Illustrative example**

<table>
<thead>
<tr>
<th>Noun, مدرسة</th>
<th>Niveau 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indefinite Noun, مدرسة</td>
<td>Niveau 1</td>
</tr>
<tr>
<td>Indefinite Nominative Noun, مدرسة</td>
<td>Niveau 2</td>
</tr>
</tbody>
</table>
## Tag set for annotation

<table>
<thead>
<tr>
<th>Tag Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>Indefinite Nominative Noun $u$</td>
</tr>
<tr>
<td>NTN</td>
<td>Indefinite Nominative Noun $un$</td>
</tr>
<tr>
<td>NND</td>
<td>Definite Nominative Noun $u$</td>
</tr>
<tr>
<td>NA</td>
<td>Indefinite Accusative Noun $a$</td>
</tr>
<tr>
<td>NTA</td>
<td>Indefinite Accusative Noun $an$</td>
</tr>
<tr>
<td>NAD</td>
<td>Definite Accusative Noun $a$</td>
</tr>
<tr>
<td>NG</td>
<td>Indefinite Genitive Noun $i$</td>
</tr>
<tr>
<td>NTG</td>
<td>Indefinite Genitive Noun $in$</td>
</tr>
<tr>
<td>NGD</td>
<td>Definite Genitive Noun $i$</td>
</tr>
</tbody>
</table>
### Experimentation and evaluation

<table>
<thead>
<tr>
<th>Morphological grammars</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb inflected form patterns</td>
<td>113</td>
</tr>
<tr>
<td>Inflected relative pronoun patterns</td>
<td>8</td>
</tr>
<tr>
<td>Broken plural patterns</td>
<td>10</td>
</tr>
<tr>
<td>Agglutination’s grammars</td>
<td>3</td>
</tr>
</tbody>
</table>
Recursive graph for parsing

Transducer for a nominal sentence

Each state of the transducer is considered as a sub-graph
... Experimentation and evaluation

Cascade for parsing

Each nominal sentence component is implemented by a separated transducer

Transducer for nominative NP
... Experimentation and evaluation

Transducer for a topic
... Experimentation and evaluation

Transducer for a nominal sentence
... Experimentation and evaluation

- The call order of transducers is fixed inspired from our carried out study
- Starting by phrases until gathering all the sentence
- Particles $\rightarrow$ Phrases $\rightarrow$ Sentences
Syntactic transducer cascade:

Total number of graphs is 50 called in a fixed order

<table>
<thead>
<tr>
<th>Order</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cas_CONJ.nog</td>
</tr>
<tr>
<td>2</td>
<td>Cas_Daref.nog</td>
</tr>
<tr>
<td>3</td>
<td>Cas_PrepZamen.nog</td>
</tr>
<tr>
<td>4</td>
<td>Cas_PrepPART.nog</td>
</tr>
<tr>
<td>5</td>
<td>CasDEM.nog</td>
</tr>
<tr>
<td>6</td>
<td>Cas_KANA.nog</td>
</tr>
<tr>
<td>7</td>
<td>Cas_INNA.nog</td>
</tr>
<tr>
<td>8</td>
<td>Cas_ProREL.nog</td>
</tr>
<tr>
<td>9</td>
<td>Cas_TOOL.nog</td>
</tr>
<tr>
<td>10</td>
<td>Cas_NG.nog</td>
</tr>
</tbody>
</table>
... Experimentation and evaluation

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>5900</td>
<td>100%</td>
</tr>
<tr>
<td>Totally disambiguated</td>
<td>4840</td>
<td>83%</td>
</tr>
<tr>
<td>Partial disambiguated</td>
<td>1060</td>
<td>18%</td>
</tr>
<tr>
<td>Failed disambiguation</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>
... Experimentation and evaluation

### Recursive Graph

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Precision</th>
<th>Recall</th>
<th>F-mesure</th>
</tr>
</thead>
<tbody>
<tr>
<td>5900 sentences</td>
<td>0.6</td>
<td>0.7</td>
<td>0.62</td>
</tr>
</tbody>
</table>

### Cascade

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Precision</th>
<th>Recall</th>
<th>F-mesure</th>
</tr>
</thead>
<tbody>
<tr>
<td>5900 sentences</td>
<td>0.74</td>
<td>0.82</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Conclusion and perspectives

Study different types of Arabic nominal sentence

Establish a transducer’s cascade to analyze Arabic corpora

Compare our new method with a recursive one

As perspectives:

Increase the granularity’s level

Generalizing our method by covering other linguistic phenomena
Thank you for your attention

Questions?