UniArab: An RRG Arabic-to-English Machine Translation Software

Dr. Brian Nolan
Yasser Salem

Introduction

- The RRG UniArab MT project
  1. An Interlingua model of Arabic MT
  2. Challenges of Arabic MT
  3. The Role and Reference Grammar (RRG)
  4. The UniArab System
- Summary and future Work
MT – Transfer approach

MT - Interlingua approach
Interlingua - considerations

- **Universal lexicon**
  - How do we construct a lexicon?
  - Must we include all distinctions made by any language?
  - How to differentiate similar items? (similar verbs in a class: 'shake vs vibrate')

- **Universal knowledge format**
  - How do we encode ‘knowledge’?
  - What should be include? (pragmatic information?)

- **Preservation of ambiguity**
An Interlingua Model of Arabic MT

- The Interlingua model incorporates:
  - source language analysis
  - universal logical structure for Arabic
  - Interlingua kernel based on the Role and Reference Grammar model

- This model is more flexible and scalable for multiple languages generation.

Challenges of Arabic MT

- Large set of morphological features.

These features are normally in the form of prefixes or suffixes that can completely change the meaning of the word.

- In Arabic there are some words that hold the meaning of a full sentence e.g., سفر، would translate to “we will travel” in English.

The free word order makes determining who did what non-trivial to a computer system.
Challenges of Arabic MT

- **Root**: a relatively invariable discontinuous bound morpheme, typically three consonants in a certain order, which interlocks with a pattern to form a stem and which has **lexical** meaning.

- **Pattern**: a bound (maybe discontinuous) morpheme consisting of one or more vowels and slots for root phonemes, which either alone or in combination with one to three derivational affixes, interlocks with a root to form a stem, and which generally has **grammatical** meaning.

- **Patterns** signify grammatical or language-internal information, distinguishing word types and classes.

- These patterns can differentiate between nouns, verbs and adjectives.

{\[
\begin{array}{|c|c|c|}
\hline
\text{Arabic} & \text{Example} & \text{POS} \\
\hline
\text{kataba} & \text{he wrote} & \text{verb} \\
\text{kataba} & \text{he corresponded} & \text{verb} \\
\text{kutib} & \text{it was written} & \text{verb} \\
\text{kutub} & \text{book} & \text{noun} \\
\text{kutub} & \text{books} & \text{noun} \\
\text{kutib} & \text{writer; (adj) writing} & \text{noun} \\
\text{kutib} & \text{writers} & \text{noun} \\
\text{mukab} & \text{desk; office} & \text{noun} \\
\text{mukitib} & \text{desks; offices} & \text{noun} \\
\text{muktabah} & \text{library} & \text{noun} \\
\hline
\end{array}
\]}

- **The root and pattern word formation** patterns in the Semitic languages have been treated by segmenting morphemes on different auto-segmental ‘tiers’ or ‘planes’ (cf. McCarthy 1982).

- Tri-consonantal **roots** which bear the basic lexical meaning of the verb occupy one tier while vocalic melodies (the **pattern**) occupy a separate tier.

- The two tiers are organised by association with a ‘template’ or ‘skeleton’ consisting of syllables of prosodic structure within words.
Challenges of Arabic MT

- Arabic has **free word order**:
  - Verb Noun Noun
  - Noun Verb Noun

- Arabic has SVO, VSO, VOS and OVS orders in sentences.

- No copula verb in Arabic; ‘to be’ or ‘to have’

Role and Reference Grammar

- RRG describes a sentence of a specific language through its:
  - Logical structure and the actor-undergoer hierarchy
  - Application of the (bidirectional) linking system and argument structure
  - Layered structure of the clause
  - Layered structure of the noun phrase
  - Layered structure of the word

- We adopt RRG as the base for a multi-language interlingua translator system.
The operator projection within the different levels of the layered structure of the clause (adapted from Van Valin 2005: 12).

(A) Verbal operators (Van Valin 2005)

**Nuclear** operators
- Aspect
- Negation
- Directionals (predicate)

**Core** operators
- Directionals (participant)
- Event quantification
- Root modality
- Negation (internal/narrow-scope)

V operators
- Tense
- Evidentials
- Illocutionary force

Role and Reference Grammar

- A simple RRG projection of an Arabic clause showing both constituents and operators within the different levels of the layered structure of the clause
Role and Reference Grammar

- Arabic allows variation in clause word order.

(a) is a SVO sentence and
(b) is the VSO equivalent.
Role and Reference Grammar

UniArab is a proof-of-concept system supporting the fundamental aspects of Arabic, such as the parts of speech, agreement and tenses.

UniArab stands for Universal Arabic machine translator system.

UniArab is based on the linking algorithm of RRG (syntax to semantics and vice versa).
The UniArab Architecture

- Phase (8) – Universal (RRG) Logical Structure Creation is the most crucial phase of the UniArab system.

- The results of the parse can be seen here with LS output of our system:

  `<TNS:PAST[do'(Khalid, [read'(Khalid,book)])]>`

`قُرِّرَ خَالِدُ الكِتَابُ qra ḥāld alktāb
الكتابَ book:N خَالِدُ Khalid:MsgN قُرِّرَ qra read:V`

- `<TNS:PAST[do'(Khalid, [read'(Khalid,book)])]>`
UniArab: RRG Arabic-English MT

Adam taught Mark mathematics.  
Omar is drinking the milk.  
James reads.  
Jack reads a lot.  
I have lost my ticket.  
I missed the plane.  
Mark is fixing the computer.  
Carl is visiting Ireland.

UniArab: RRG Arabic-English MT
UniArab: RRG Arabic-English MT

Processing unrecognised Arabic words

In a case:

1. where a word is not available in the lexicon
2. but the logic structure is recognised

UniArab will output a correctly structured translation, but with the unknown Arabic word in its position within the English sentence.

• This makes the system resilient to slight misspellings which can be recognised and corrected.
UniArab’s Lexicon Interface

The graphical interface is quicker and easier when a user adds a new word in the lexical (XML data source).

UniArab – testing and evaluation

- Test strategy
  - verb-subject agreement
  - Demonstrative adjective-noun agreement
  - Gender-ambiguous proper nouns
  - ‘to be’ and ‘to have’
  - Free word order (Verb Noun Noun)
  - Pro-drop

<table>
<thead>
<tr>
<th>Arabic</th>
<th>human-translated</th>
</tr>
</thead>
<tbody>
<tr>
<td>قراء جاد الكتاب</td>
<td>I read the book.</td>
</tr>
<tr>
<td>قراء ماري الكتاب</td>
<td>Mary reads the book.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arabic</th>
<th>human-translated</th>
</tr>
</thead>
<tbody>
<tr>
<td>يشرب عمر الهليب</td>
<td>Omar is drinking milk.</td>
</tr>
<tr>
<td>شرب عمر الهليب</td>
<td>Omar drank the milk.</td>
</tr>
<tr>
<td>مارك قرأ الكتاب</td>
<td>Mark reads the book.</td>
</tr>
<tr>
<td>سيشرب مارك اللبن</td>
<td>Mark will drink the milk</td>
</tr>
</tbody>
</table>
UniArab – testing and evaluation

RRG Lexicon in XML

Verb lexicon - fragment as XML

Noun lexicon - fragment as XML
RRG Lexicon in XML

```xml
<process name="read">
  <subject>John</subject>
  <object>
    <read>
    </read>
  </object>
</process>
```

Java code fragment:

```java
public static String verbGender = "M";
public static String LSI = "";
public static String LSB = "";
//public String[] allAttribute2;
public static void generateLSI(String allAttribute, String[] ArabicSentence) {
  //who do what
  for(int n = 0 ; n < allAttribute.length; n++) {
    if (allAttribute[n].equals("Verb")) {
      //there is a verb
      if (allAttribute[n+1].equals("N")) {
        //the gender of verb = M
        verbGender = "M";
        //Add LS to string LS
        LSI = allAttribute[n+2];
      } else if (allAttribute[n+1].equals("F")) {
        //the gender of verb = F
        verbGender = "F";
        LSI = allAttribute[n+2];
      } else if (allAttribute[n+1].equals("NoGender")) {
        //the gender of verb = F
        verbGender = "NoGender";
        LSB = allAttribute[n+2];
      }
    }
  }
  //and loop who do what
```

This fragment determines appropriate gender marking.
Summary and Future work

- We have discussed our rule-based lexical framework for Arabic language processing using RRG.

- The framework is supported by a system called UniArab.

- UniArab is an RRG motivated XML-based machine translator implemented using Java.

- Our framework and system demonstrates the capabilities of RRG as a basis for multi-language interlingua machine translation systems.

Summary and Future work

Our main future work:

- **Supporting** more complex Arabic sentences and word ambiguity.

- **Comparing** UniArab against MT systems based on non-RRG techniques, such as systems provided by Google and Microsoft.

- **Evaluating** the performance of UniArab with multi-languages processing.

- **Adding** extra languages with interlingua bridges.

- **Connecting** the lexicon with a rich concept ontology to enhance the granularity of the logical structures.
Linguistics and KB Interlingua

**Linguistics-Based Interlingua:**
- **Linguistic model** (+ its semantic and linking theory) identifies meaning representations and constraints and hence provide a sufficient basis for an interlingua representation.

**Knowledge-Based Interlingua:**
- **Linguistic meaning** is dependent on non-linguistic knowledge.
- Use **real world knowledge** to augment meaning representations.
Thank you ……… !

Questions?

brian.nolan@itb.ie