A Proposed Approach for Generating Arabic from Interlingua in a Multilingual Machine Translation System

Azza Abdel Monem
Central Lab. For Agricultural Expert Systems (CLAES),
P.O. Box: 100 Dokki, Giza, Egypt.
E-mail: azza@mail.claes.sci.eg

Khaled Shaalan
Computer Science Dept., Faculty of Computers and Information
Cairo Univ., 5 Tharwat St., Orman, Giza, Egypt.
E-mail: shaalan@mail.claes.sci.eg

Ahmed Rafea
Computer Science Dept., American University in Cairo
113, Sharia Kasr El-Aini, P.O. Box 2511, 11511, Cairo, Egypt.
E-mail:rafea@aucegypt.edu

Hoda Baraka
Computer Engineering Dept., Faculty of Engineering, Cairo University,
Dokki, Giza, Egypt
Hbarka@idsc1.gov.eg

Abstract

Intelingua (meaning) representation has been successfully used in multilingual machine translation. This paper reports our attempt to generate Arabic sentence from interlingua. The proposed system will be compatible with the NESPOLE[1] consortium. In NESPOLE an Interlingua called interchange format or IF, designed for travel planning is used. Our approach describes how to generate grammatically correct Arabic sentence from Interlingua. It involves two main components a mapper for converting intelingua into syntactic structure (feature-structure) and a generator for generating the target Arabic sentence that represents the intended meaning. A translation example is provided to explain the inner working of the system.

1. Introduction

This paper describes the integration of an Arabic generation system, developed in the framework of a collaborative research project[2] entitled "Machine Translation of Spoken Arabic into English" between Cairo University and Carnegie Mellon University, with the NESPOLE consortium's interlingua-based machine translation. The system will translate simple conversations between a travel agent speaking Arabic, and a traveler speaking American English.

[1] NESPOLE— NEgotiating through SPOken Language in E-commerce. See the project web site at http://nespole.itc.it

[2] NSF Project Number: INT-0001613 US Egypt Joint Science and Technology Board Project Number: INF4001023.

Interlingua-based MT has a number of advantages over other approaches, such as the 'transfer' model. In an Interlingua-based architecture, source text analysis and target text generation are divided into separate components. A language-independent intermediate representation (or Interlingua) mediates between these two components. The decoupling of the analysis and generation phases allows the system to handle multiple-language output and avoids the reconfiguration of the system for each new language [10] [12].

The NESPOLE system is an Interlingual approach to machine translation [8] [11]. In NESPOLE an Interlingua called *interchange format* or *IF*, designed for travel planning is used (e.g making hotel and flight reservations, etc.) [6] [7]. The NESPOLE interlingua is based on domain actions, which in our case are things you can do when you are talking about a trip. Domain actions include requesting information about availability of hotel rooms, giving information about the price of flights, requesting a reservation at a hotel etc. There are six languages in NESPOLE, English, French, Italian, German, Japanese, and Korean. Arabic will be the seventh in this family. Aim of this research work will add the Arabic generation system to the NESPOLE project. This will make it possible to translate for six languages into Arabic.

Our approach describes how to generate grammatically correct Arabic sentence from Interlingua. It involves two main components a mapper for converting intelingua into syntactic structure (feature-structure) and a generator for generating the target Arabic sentence that represents the intended meaning.

The rest of this paper is structured as follows. Section 2 describes the system that generates Arabic sentences from interlingua representation and show how basic sentential components are mapped. Section 3 gives a description of the interlingua representation. Section 4 and 5 describes the main components of the generation system: the mapper module and the generator module. In this context, we address some issues in generating Arabic from interlingua such as agreement in number which cannot be transferred exactly from the IF of an input sentence, e.g. English sentence. An example translation are provided in section 6. In this context, we address the issue of word order variation in Arabic. In Section 7, we give some concluding remarks.

2. The Architecture of the Arabic Generation System

NESPOLE is an international consortium for translation of spoken language dialogues about travel planning (e.g making hotel and flight reservations, etc.)[6] [7]. In this system, each sentence is first conveyed into tokens. The system analyzer uses a lexicon, a morphological analyzer, source language grammar and semantic information in order to parse the tokenized sentence into a feature structure (FS), a list of feature-value pairs that reflects the syntactic structure of the source language (e.g., English). The interpreter then uses mapping rules to convert the FS into an Interlingua. An Interlingua is a tree-structured representation that abstracts away many of the syntactic details of both source and target language, while conveying the meaning of the source language.

Generation of the target language sentence begins with the Interlingua [18]. The developed system takes Interlingua of a sentence as input and constructs the syntactic structure of the target sentence as output by utilizing various knowledge recourse fed into the system. The basic architecture of proposed system is shown in Figure 1. The system consists of two main components. The first component is the *Mapper* and the second component is the *Generator*.

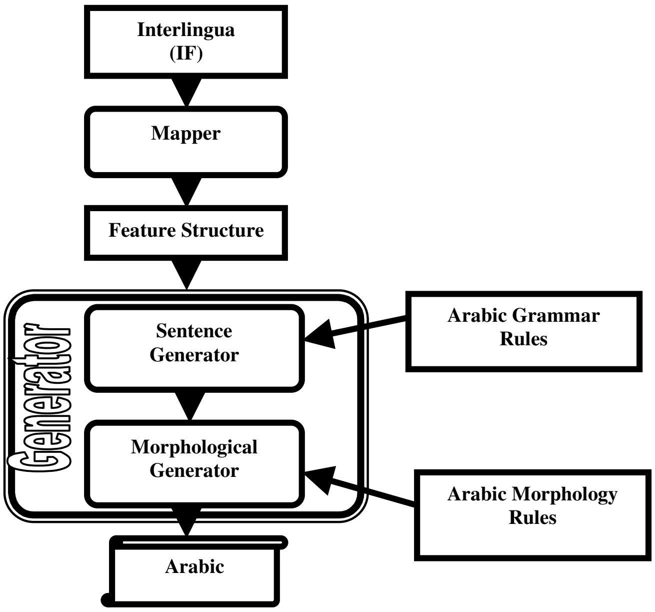


Figure 1: The Architecture of the Arabic Generation System

First, the generation mapping rules convert the Interlingua into an FS that reflects the syntactic structure of the target language. An FS is a list of feature-value pairs that reflects the syntactic structure of the target language. Target language lexicon entries are FSs. They are retrieved during mapping and added to the sentence FS under construction [15].

3. The Description of Interlingua

We are working on travel planning, a task-oriented domain. In a task-oriented domain, The Interlingua representation known as the Interchange Format (IF) [9]. IF is based on a set of domain actions (DA) with parametric arguments. Each DA has up to four components: the *speech act*, the *concepts*, the *arguments*, and a *speaker tag*. Plus sign separate speech acts from the concepts and concepts from each other. In general, each DA has a speaker tag and at least one speech act optionally followed by string of concepts and optionally, a string of arguments. DAs can be roughly characterized as follows:

$$\underbrace{\text{Speaker:}}\ \underbrace{\text{speech act}} + \underbrace{\text{concept}^*}\ \underbrace{\text{arguments}^*}$$

```
(1)   a:on the twelfth we have a single and a
      double available.
      a: give-information+availability+room
      (provider=we, room-
      type=(operator=conjunct, [(single_room,
      quantity=plural),  (double_room,
      quantity=plural)], time=(md=12))

(2)   a:and we+ll see you on February twelfth.
```

```
       a: greeting (conjunction=discourse,
       greeting=goodbye, to-whom=we,
       time=(month=2,md=12))

(3)    c:thank you very much
       c: thank
```

In example (1) the speech act is give-information, the concepts are availability and room and the arguments are time and room-type. The possible arguments of DA are determined by inheritance through a hierarchy of speech acts and concepts. In this case time is an argument of availability and room-type is an argument of room. Example (2) shows a DA which consists of speech act with no concepts attached to it. The argument time is inherited from the speech act greeting. Finally, Example (3) demonstrates a case of DA which contains neither concepts nor arguments. The following paragraphs describe the four components of DAs, speaker tags, speech acts, concepts, and arguments.

Speaker tag: The speaker tag is either a: for agent or c: for customer to indicate who is speaking. The speaker tag is sometimes the only difference between the IFs of two sentences. For example, "Do you take credit card?" (Uttered by the customer) and "Will you be paying with credit card?" (uttered by the agent) are both requests for information about the credit cards as a form of payment.

Speech Act: Some speech acts are very general. For example, give-information is used in many DAs where the speaker intent is to inform the listener of something, such as give-information+expiration-date, etc. In most cases, each representation contains a simple speech act. But verification and negation acts are the exception and can combine with other speech acts and form complex one. For example the sentence "So you're not leaving on Friday, right?" has the speech act request-verification-negate-give-information.

Concepts: Each DA can have zero or more concepts following the speech act, although not all possible strings of concepts are allowed. Concepts fall into several classes that roughly constrain how they combine with each other. Some classes of concepts are actions (change, reservation, confirmation, cancellation, etc.), attributes (availability, size, temporal, price, location, features, etc.), and entities (room, hotel, expiration date, etc.).

The usual order of concepts in a DA is action+attribute+entity as in request-action+reservation+temporal+room for " I'd like to make a reservation for a room on the fifth.

Arguments: Arguments add specific information to the DA, such as times, prices, and specific features of entities. An argument consists of an argument name and a value separated by an equal sign, for example room-type=double

4. The Architecture of Mapper Module

This is the first module of the proposed system that performs the first task, FS creation in a language independent way. The developed system takes the IF of a sentence as input and constructs the FS of the target language as output by utilizing various knowledge resources fed into the system. In other words, the system makes transfer between two representation languages, IF (interlingua representation) and FS (feature structure) representation. To achieve this task, two knowledge resources are utilized by the system: *ontology* and *lexicon*. The architecture of mapper is shown in Figure 2.

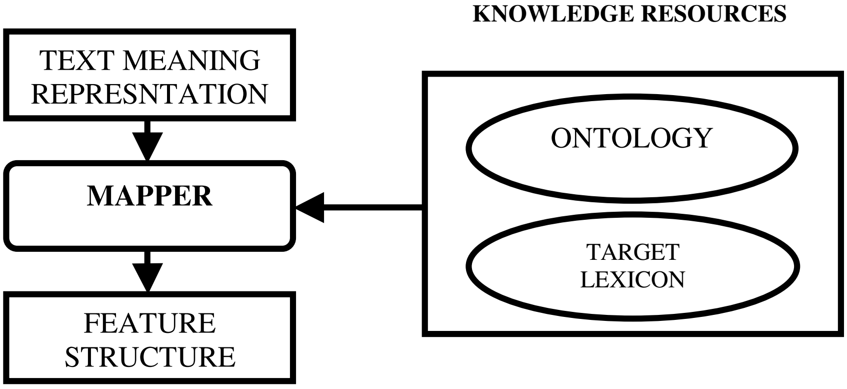
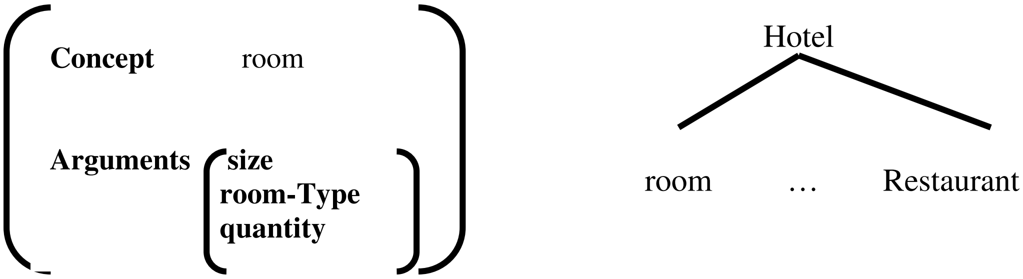


Figure 2: Components of the Mapper Module

Ontology: a hierarchical representation of speaker's world knowledge about entities, events, and their relationships. Every entry, called a *concept*, in the ontology is a primitive symbol that represents a proposed abstraction about a set of things in the world. Each concept represents either a group of entities or a set of similar events. A concept which is created for a group of entities decomposes the definition into a set of arguments that any entity from that group can take it. Each argument takes its values from a well-defined domain and representation of real entities are achieved through instantiating these properties with specific values. See Figure 3.



3.a A concept specification **3.b** A concept Hierarchy

Figure 3: Examples of Ontology

Lexicon: a bilingual dictionary for mapping between English and Arabic. All the words of the languages are defined in the lexicon. It contains information about the concepts, arguments, and values. Each entry in the lexicon corresponds to a word sense of the target language and provides information about word's phonological, morphological, syntactic, and pragmatic properties. Such information can be used in the selection of words to be used in target sentences. Each entry consists of a number of features. Examples of the features which can be used in the definition are: *CAT* (syntactic category, such as verb, noun), *Gend* (Gender, such as male, female), *NUM* (number, such as singular, plural), PER (personal, such as 1st person, 2nd person), *TENS* (tense, such as past, present), and TRANS (transitivity, such as transitive, intransitive). To exemplify how word is defined in the lexicon, the following example is given.

Room "غرفة" (ghurfah)	
CAT	noun
GEN	female
NUM	singular

Desire "أرغب" (arghub)	
CAT	verb
TENSE	present
TRANS	transitive
GEN	neutral
NUM	singular

Feature Structure (FS): It is the output of mapper which represents the interface between interchange format (IF) and the target sentence. This is the simplest ways of to encode the kinds of properties that we have in mind is through the use of feature structure.

5. The Architecture of Generator Module

This is the second module of the proposed system. The developed system takes the feature structure of a sentence as input and constructs the target Arabic sentence as output by using both Arabic grammar rules and morphology rules that make the generation of Arabic sentence from the feature structure grammatical correct and readable.

5.1 Arabic Grammar Rules

A set of Arabic grammar rules is applied to generate the Arabic sentence in the target language to get a grammatically correct structure [14] [13]. These Arabic grammar rules are divided into many categories based on the syntactic structure of Arabic sentence. The syntactic structure of Arabic sentence can be analyzed into two main constructional categories: nominal sentence and verbal sentence [17]. The nominal sentence consists of inchoative "المبتدأ" and enunciative "الخبر". There are some special cases for the nominal sentence. The following are some examples for grammar rules of nominal sentence:

1. *inchoative "المبتدأ" should be in nominative case ("مرفوع")*
2. *A quasi-sentence ("شبه الجمله") starts with a preposition ("حرف جر").*
3. *the postfixed noun ("المضاف الية") is Defined*

The verbal sentence consists of a verb "فعل", subject "فاعل أو نائب فاعل" and object "مفعول به". The following some examples for grammar rules of verbal sentence:

1. *A verbal sentence starts with a verb.*
2. *The transitive verb refers to an object.*
3. *There should be an agreement between the verb and the subject in gender.*

5.2 Arabic Morphology Rules

The Arabic morphology will generate the inflected Arabic word according to the Arabic conjugation and agreement rules [1] [3]. This relationship is between words in certain context such that a word in one position follows the word in a corresponding position in some aspects: such as number (single, plural), gender (male, female), and definition (definite, indefinite) [2] [5].

5.3 Issues in Generating Arabic from Interlingua

We consider some issues in generation of correct agreement in Arabic sentences from an Interlingua (IF) between the source language (English) and the target language (Arabic) [4].

(i) Subject-Verb(SV)/Verb-Subject(VS) Agreement: In Arabic, agreement in number between subject and verb depends on the nature of the subject of the sentence and word order. On a VS order, verbs do not agree in number with a plural subject. Agreement is always singular. Verbs, however, agree with their subjects in person and gender [16]. For example, the "the kids would like to visit pyramids" has the following translation according to the verb and subject order:

- SV order: الأولاد يرغبون في زيارة الأهرام
  (alawlad yarghaboona fi ziyaratu alahram)
- VS order: يرغب الأولاد في زيارة الأهرام
  (yarghabu alawlad fi ziyaratu alahram)

(ii) Intrinsic Number: In most cases, the number feature for a noun is determined by the input sentence, reflected in the IF, and mapped directly from the IF into the FS by the mapper. Some nouns, however, may have agreement constraints already present in the lexicon. While lexical entries for nouns are usually assumed to be singular, certain nouns may be intrinsically plural in terms of agreement. For example, the noun ناس *(nas)* 'people', would contain the agreement information *(NUM plural)* in the lexicon, and the mapper should not override it with information that may be present in the Interlingua (for example, if the source language were Italian or Spanish, in which the word is a singular collective noun).

(iii) Number-Noun Agreement: Number-noun agreement is governed by a set of complex rules. With the number 'one', agreement is as expected, but there may be a reversal of word order (e.g. رجل واحد *(rajulun waaHidun)* 'one book (مرفوع nominative)'). The number 'two' is expressed by the dual of the noun. Numbers 'three' through 'ten' require the noun to be plural and the gender of the number to be the opposite of the gender of the singular noun. For example: خمس (*khams)* 'five' (masculine) سنوات (*sanawaat)* (plural of سنة *sanat* 'year', feminine) but خمسة *(khamsatu)* 'five' (feminine) متاحف *(matahif*) (plural of متحف (*mathaf)* 'museum', masculine). Up to ten (plural of paucity), numbers and nouns agree in case, which is determined by the syntactic construction they appear in. Numbers above ten (plural of multiplicity) require a singular noun in the indefinite accusative. Agreement decisions can be made in the generator with the help of a callout function, but are most easily handled using the mapper.

6. An Example Translation and Results

To demonstrate the function of the components described in the sections above, we will use the English-Arabic translation example below:

The input English sentence

```
c: I would like to reserve a single room
```

Assume that the English NESPOLE analyzer generates the following IF:

The Interlingua (IF)

```
c:give-information+disposition+reservation+room
(disposition=(desire,   who=i),   room-spec=(single-room,
quantity=1))
```

In the current system, the mapper takes as input this IF and produces the FS for Arabic:

The feature Structure (FS)

```
[c:,[give-information],[disposition],[reservation,   Cat
noun,   "حجز",   Num   sg,   Gen   male],[room]
(disposition=([desire, verb, "رغب", Num sg, Gen male,
Tense past], who=[I, Cat pronoun, "أنا", Person:
1ST, Num sg]), room-spec=([single-room, Cat adj, "فردي
", Num sg, Gen male], [Cat noun, "غرفة", Num sg,
Gen female, Irr_PL "غرف"]))
```

Note that Most of the linguistic features used in the in the FS (e.g., tense, argument class, number, person) should be self-evident. The resulting FS serves as input to the Arabic morphological and sentence generator, producing the Arabic surface form:

The output Arabic sentence

أنا أرغب في حجز غرفة فرديه (ana arghabu fi hagzi ghurfatun fardiyah)

A major problem with the current implementation of the system relates to the word order variation in Arabic. Arabic is basically a VSO language, in which constituents can change order according to the constraints of text flow or discourse. The grammatical roles of constituents are identified by explicit morphological case markings. For example, there is no information structure for the system to decide whether to generate a VS order or an SV order from an IF for the English sentence. Currently, the system produces all sentences in the SV order.

7. Conclusion

While there are challenges to be worked out where the source language and target language differ greatly in their morphology and syntax, an Interlingua approach allows for a flexible integration of software modules for languages that differ in their realization of the same unit of meaning. Indeed, most of the morphological and syntactic differences between the source language and the target language can be handled by either the mapper or the generation grammar.

In this paper, we have described an ongoing research project for integrating an Arabic generation system with an interlingua-based machine translation system. The proposed system will be compatible with the NESPOLE consortium. NESPOLE is a multinational research project for supporting multilingual machine translation. The system translates simple conversations between a travel agent speaking Arabic, and a traveler speaking American English. After giving a description of the system that generates Arabic sentences from IFs, we have shown how basic sentential components are mapped. In this context, we have addressed some of the problems faced in the translation between English and Arabic, such as agreement in number which cannot be transferred exactly from the IF of an English sentence. We have also provided an example translation and results.

References

1. Beesley, K. (1996). Arabic Finite-State Morphological Analysis and Generation. In *the Proceedings of the COLING'96*, Vol. 1, pp. 89-94.
2. Cavalli-Sforza, V., Soudi, A. and Mitamura, T. (2000) Arabic Morphology Generation Using a Concatenative Strategy. In *the Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, Seattle, April 29-May 3, pp. 86-93.
3. Cavalli-Sforza, V., Soudi,A. (2003). Enhancements to a Morphological Generator Motivated by English-to-Arabic MT. *In the Proceedings of the Eight International Symposium on Social Communication*, Center for Applied Linguistics, SANTIAGO DE CUBA, January 20-24.
4. Fehri, F., (1993). *Issues in the Structure of Arabic Clauses and Words*. Kluwer Academic Publishers, Dordrecht, Holland.
5. Leavitt, J.R., (1994). *MORPHE: A Morphological Rule Compiler.* Technical Report, CMU-CMT-94-MEMO.
6. Lavie A.., Fabio P., and Loredana T. (2001). Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications. *the Human Language Technology Meeting (HLT-2001)*, San Diego, March.
7. Lavie A., Levin L., Schult T., and Waibel A. (2001) Domain Portability in Speech-to-Speech Translation. *In the Proceedings of the Human Language Technology Meeting (HLT-2001)*, San Diego, March.
8. Lavie, A.; Waibel, A.; Levin, L. (1997). Janus III: speech-to-speech translation in multiple languages. *In the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Munich 1997. Los Alamitos, Calif. 1997. Universit?t Karlsruhe; Institut für Logik, Komplexit?t und Deduktionssysteme.
9. Levin L., Gate D., Lavie A., and Waibel A. (1998). An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues, *In the Proceedings of International Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Australia, November.
10. Nirenburg S., Carbonell J., Tomita M., and Goodman K. (1992). *Machine Translation: A Knowledge-based Approach.* Morggan Kaufmann, San Mateo, California.
11. Mitamura, T., Nyberg, E.H. and Carbonell, J. (1991). An Efficient Interlingua Translation System For Multilingual Document Production. *In the Proceedings of the 3rd Machine Translation Summit.*
12. Nyberg, E.H. and Mitamura, T. (1992). The KANT System: Fast, Accurate, High Quality Translation in Practical Domains, *In the Proceedings of COLING'92.*
13. Soudi, A., Cavalli-Sforza,V. and Jamari, A. (2001): A Computational Lexeme-Based Treatment of Arabic Morphology. *In the Proceedings of the Association for Computational Linguistics Workshop: Arabic Language Processing: Status and Prospects*, Toulouse, France, July, pp. 155-162.

14. Soudi, A., Cavalli-Sforza,V. and Jamari, A. (2002a): Arabic Noun System Generation. *In the Proceedings of The International Conference on the Processing of Arabic*, Lamanouba University, Tunisia, April, pp. 69-87.
15. Soudi, A., Cavalli-Sforza,V. and Jamari, A. (2002b): A Prototype English-to-Arabic Interlingua-based Machine Translation System. *In the Proceedings of The Arabic Language Resources (LR) and Evaluation: Status and Prospects workshop, Third International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Spain, June 1.
16. Soudi, A., Cavalli-Sforza,V. (2002c): Arabic Morphology Generation: A two-step strategy. *In the Proceedings of the Arabic and Information Technology International Conference*, organized by le Haut Conseil de la Langue Arabe, Algiers, Algeria, 28-29.
17. Timothy, A.B. (1990). Lexicographic Notation of Arabic Noun Pattern Morphemes and their Inflectional Features, *In the Proceedings of the Second Cambridge Conference on Bilingual Computing in Arabic and English*. No pagination.
18. Wilcock G. and Jokinen K. (2003). Generation Models for Spoken Dialogues, 2003 AAAI Spring Symposium, Stanford. *In Natural Language Generation in Spoken and Written Dialogue*, Technical Report SS-03-06, American Association for Artificial Intelligence, pp. 159-165.