Challenges in the Generation of Arabic from Interlingua

Abdelhadi Soudi

CLC-Ecole Nationale de L'Industrie Minérale, Av. Hadj Ahmed Cherkaoui, B-P: 753 Agdal, Rabat, Morocco

asoudi@enim.ac.ma

Résumé-Abstract

Nous décrivons un ensemble de problèmes cruciaux que nous avons rencontrés au niveau de la génération des phrases de la langue Arabe à partir des représentations interlangues utilisées dans le système KANT « Knowledge-based Accurate NL Translation », un système de traduction automatique basé sur des représentations interlangues qui permet de traduire vers de multiples langues (Nyberg, E.H. and Mitamura, T. (1992)). Les problèmes que nous avons rencontrés sont catégorisés en deux groupes : les problèmes liés aux différences linguistiques entre l'Anglais et l'Arabe et les problèmes liés au système de transformation « mapper ». Ces problèmes ont un effet non seulement sur la qualité de la traduction mais aussi sur la grammaticalité des phrases générées. Nous montrons comment nous avons traité ces problèmes. Les résultats expérimentaux sont aussi présentés.

We describe a set of crucial problems we have encountered in the generation of Arabic sentences from the Interlingua representations used in the KANT knowledge-based machine translation (MT) system, an interlingua-based software architecture for translation from English to several languages (Nyberg, E.H. and Mitamura, T. (1992)). These problems are categorized into two groups : problems related to the language mismatches between English and Arabic and problems related to the mapping system. The major language mismatches have to do with verb classes and alternations in the two languages, word order, tense and agreement differences. With respect to the issues related to the mapping system, the mapping system does not provide, inter alia, any information structure that would accommodate word order variation in Arabic. These problems have great influence not only on the quality of the translation but also on the acceptability of the sentence generated. We show how we have handled some of these problems. Experimental results concerning English-to-Arabic MT are also presented.

Mots Clés-Key Words:

Traduction automatique, génération, Arabe, Anglais, Représentations Interlangues.

Machine Translation, generation, Arabic, English, Interlingua Represenations.

1. The Arabic Generation System

Generation of the target language sentence begins with the Interlingua Representation. The system which generates Arabic sentences from Interlingua Representations consists of 4 subsystems: the mapping system, the sentence generation system, the sentence/morphology generation interface and the morphological generation system, as shown in Figure 1 below:

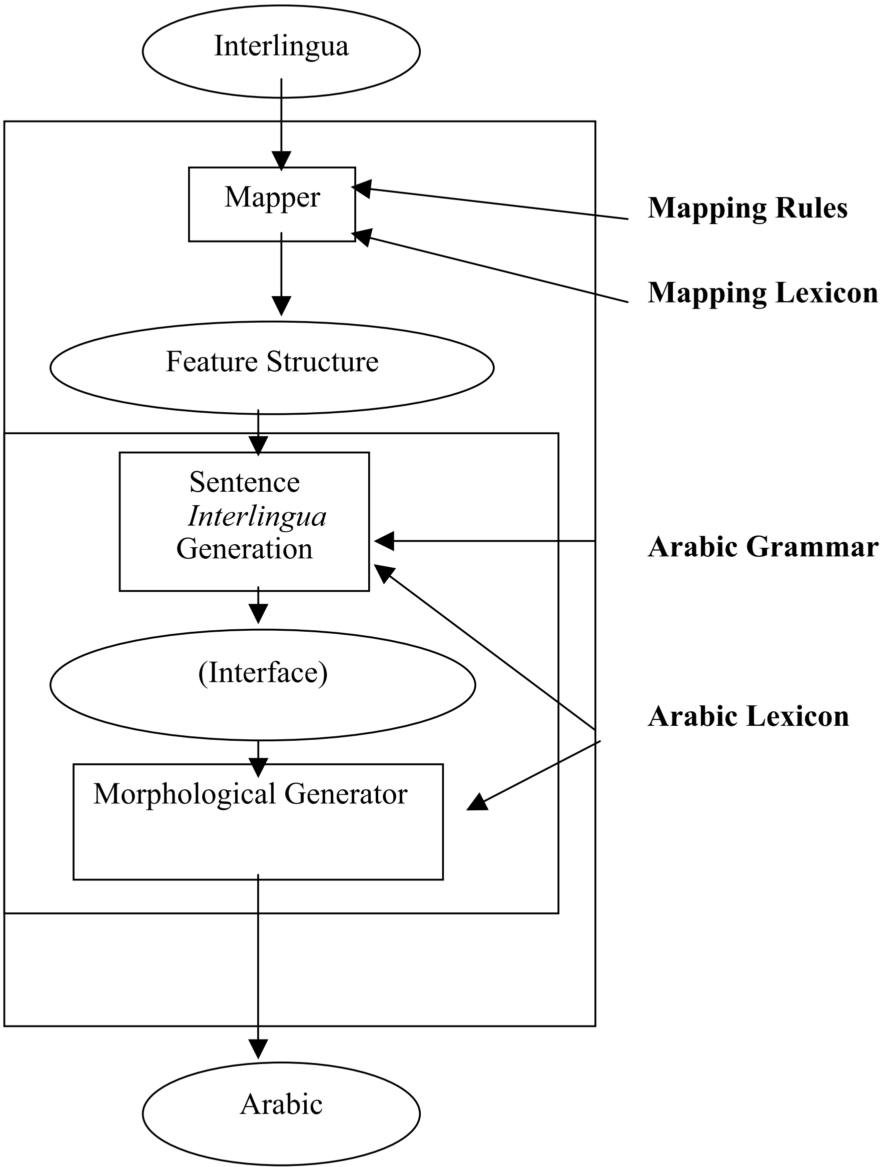

Figure 1: The Arabic Generation System[1]

[1] *The morphology/generation interface consists of a lisp program that defines some functions that are used to call the morphological generator from the sentence generator.*

To demonstrate the function of these subsystems, we will use the example sentence below:

(1) Jakarta and Bangkok are shining the most.

The KANT analyzer generates the following Interlingua Representation for the sentence in (1) :

(2)

The Interlingua

```
(*A-SHINE (FORM FINITE) (TENSE PRESENT) (MOOD DECLARATIVE)
(PUNCTUATION PERIOD) (PROGRESSIVE +) (IMPERSONAL -) (ARGUMENT-
CLASS AGENT) (MANNER (*M-THE-MOST (POSITION POSTVERBAL) (UNIT -)
(DEGREE POSITIVE))) (AGENT (*G-COORDINATION (PERSON THIRD) (IMPLIED-
REFERENCE +) (CONJUNCTION (*CONJ-AND)) (CONJUNCTS (:MULTIPLE (*PN-
JAKARTA (UNIT -) (PERSON THIRD) (NUMBER SINGULAR) (REFERENCE NO-
REFERENCE)) (*PROP-BANGKOK (UNIT -) (PERSON THIRD) (NUMBER SINGULAR)
(REFERENCE NO-REFERENCE)))))))
```

Most of the linguistic features used in the KANT Interlingua and Feature structure (FS) (e.g., punctuation, form, tense, argument class, number, person) should be self-evident. Some other features are artifacts of KANT's evolution as a technical text system. The *IMPLIED-REFERENCE* feature is used for nouns, such as the proper noun in the example above. *G-COORDINATION* contains all conjuncts that are coordinated and the conjunction that is used.[2] It is beyond the scope of this paper to present a detailed description of the software modules of the KANT system.

In the current system, the mapper takes as input the Interlingua Representation in (2) and produces the FS for Arabic (3), using a set of mapping rules and a mapping lexicon. 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.[3]

(3)

```
((ADV ((CAT ADV) (ROOT "?akθar"))) (form 4)  (CAT V) (ROOT "ta?allaq") (VOICE
ACT) (TENSE IMPERF) (MOOD INDIC) (SUBJ ((ELEMENT (*MULTIPLE* ((AGR
```

[2] *To promote representational consistency, the same structure is (*G-COORDINATION*) is used if there is no explicit conjunction. In this case, the feature* CONJUNCTION *will have the value* NULL.

```
((GENDER F) (PERSON 3) (NUMBER SG))) (CAT N) (ROOT "jakarTaa")) ((AGR
((GENDER F) (PERSON 3) (NUMBER SG))) (CAT N) (ROOT "baankuuk")))) (CONJ
((CAT CONJ) (ROOT "wa"))))) (PUNCTUATION ((ROOT PERIOD))))
```

The resulting FS serves as input to the Arabic morphological and sentence generator, producing Arabic surface forms:

(4)

```
baAnkuwk wa jakaroTaA tata^alGaqaAni ^ako#ar
```

2. Issues in the generation of Arabic from an Interlingua Representation

In this section, we describe a set of crucial problems we have encountered in the generation of Arabic sentences from an Interlingua Representation and show how we have handled some of them. These problems have great influence not only on the quality of the translation but also on the acceptability of the sentence generated.

2.1. Language Mismatches:

The major language mismatches have to do with word order, agreement and tense differences. Additionally, verb classes and alternations in the two languages are not always the same.

To generate Arabic sentences, we have used Genkit (Generation) Kit (Tomita M. and Nyberg E.H. (1988)), a system that compiles a grammar written in a formalism called Pseudo-Unification Grammar into a sentence generation program. The generator follows a top-down, depth-first strategy for applying rules during generation.

The following example shows a unification-based grammar rule for generating sentences. The rule consists of a context-free phrase structure rule and a list of pseudo equations :

(5)

```
(<S> ==> (<NP> <VP>)

        ((x1 agr) = (x2 agr))

        (x1 == (x0 subj))
        (x1 case) = nom)

        (x2 = x0)
```

The non-terminals in the phrase structure part of the rule are referenced in the constraint equations as x_0 ... x_n, where x_0 is the non-terminal in the left-hand side (here, <S>) and x_n is the *n-th* non-terminal in the right hand side. In these equations, x_1 represents <NP> and x_2 represents *<VP>*. The rule in (5) is for sentences with an *<NP>* and a *<VP>* that agree in number, person and gender. The equation *((x_1 agr) = (x_s agr))* indicates that the *<NP>*'s agr feature has a value that unifies with the value of the *<VP>*'s agr.

The generation of properly inflected Arabic verbs and nouns is a concern of both the mapper and the generator. For example, the generation of correct agreement between nouns and their modifiers or other parts of the sentence may be performed either during mapping or during generation. Different cases must be considered:

(i) Subject-Verb/Verb-Subject Agreement: In Arabic, contrary to English, 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, as is illustrated by the following rule for generating a VS order sentence:

(6)

```
(<s> ==> (<vp> <np>)

    (((x1 agr) = (x0 subj agr))

     ((x1 agr number) <= 'sg)

     (x2 == (x0 subj))

     ((x2 case) = nom)

     (x1 = x0)))
```

In this rule, the overwrite assignment equation *((x1 agr number) <= 'sg)* means that the old value of the verb number feature should be overwritten by the singular. For example, if the verb is assigned the value plural for the number feature in the Interlingua Representation, then this value should be replaced by the singular.

(ii) Intrinsic Number: In most cases, the number feature for a noun is determined by the input sentence, reflected in the Interlingua Representation, and mapped directly from the Interlingua Representation 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 *naAs* ‘people’, would contain the agreement information *(number pl)* 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).

In the case of sound plural feminine nouns, such as *Hayawanaat* “animals”, agreement in number and gender is singular and feminine, respectively regardless of the values of these features present in the Interlingua Representation. This issue is handled using the following mapping rule:

(7)

```
(:test (:sem (number plural)
      :syn (:not (human +)))

 :force-add ((agr ((gender f) (number sg)))))
```

The mapping rule above consists of a set of slots and values associated with the noun mapping hierarchy node. The *:TEST* slot specifies a set of conditions that must be passed for the rule to be applied. The *:SYN* subslot specifies a negated condition on the FS, namely the feature *(:not (human +)*, that must be met. The *:SEM* subslot specifies a condition on the Interlingua Representation, namely the FVP (number plural). The slot *:force-add* indicates that the FS under construction should have feminine as its gender value and singular as its number value. This slot actually overrides information in the Interlingua Representation: the value of the number feature in the Interlingua Representation, namely plural, is overridden here by the singular. The mapping rule above applies to the sound plural feminine in Arabic (i.e., the *-At* class). By way of example, in the Interlingua Representation for the noun "animals", we would have, inter alia, the feature-value *pair (number plural)*. This information should be overridden for the corresponding Arabic noun -which is (human -)- by the feature-value pairs (number singular) and (gender feminine). Note that the information specified by the *:force-add* slot in the example above relates to subject-verb agreement.This slots simply adds the value feminine for the gender feature. Thus, the sound plural noun *Hayawanaat* is plural but has 'singulative' agreement with verbs.

(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. *kitaabun waaHidun* ‘one book (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: *xams* ‘five’ (masculine) *sanawaat* (plural of *sanat* ‘year’, feminine) but *xamsatu* ‘five’ (feminine) *kutub* (plural of *kitaab* ‘book’, 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.

2.2. Issues Related to the mapping System

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. However, the KANT analyzer does not mark constituents as topic or focus. That is, the mapping system does not provide such information structure in the Interlingua Representation. For example, there is no information structure for the system to decide whether to generate a VS order (9a) or an SV order (9b) from an Interlingua Representation for the English sentence in (8).

(8) Zayd ate the apple.

(9)

a. ?akala zayd-un t-tuffaaHat-a.

ate Zayd-nom the-apple-acc

b. zayd-un ?akala t-tuffaaHata

Zayd-nom ate the-apple.

Currently, the system produces all sentences in the S(=topic)V order.

3. Results

The system has been able to translate completely and correctly 30 sentences of 41 sentences in the domain of broadcast news captions. The system has been tested on different structures and has produced good results. The system is still under construction. The coverage and accuracy of the system can be further extended by adding more generation (morphology and sentence) and mapping rules.

Conclusion

In this paper, we have described some problems we have encountered in the generation of Arabic sentences from the Interlingua Representation used in the KANT knowledge-based MT system, an interlingua-based software architecture for translation from English to several languages. We have categorized these problems into two groups : problems related to the language mismatches between English and Arabic and problems related to the mapping system. The major language mismatches have to do with verb classes and alternations in the two languages, word order, tense and agreement differences. With respect to the issues related to the mapping system, the mapping system does not provide, inter alia, any information structure that would accommodate word order variation in Arabic. We have also provided an example translation and results.

References

Aronoff, M. (1994), *Morphology by Itself: Stems and Inflectional Classes*, Cambridge, MIT Press.

Beard, R. (1995), *Lexeme-Morpheme Base Morphology: A General Theory of Inflection and Word Formation*, State University of New York Press.

Cavalli-sforza, V., Soudi, A., Mitamura, T. (2000), Arabic Morphology Generation Using a Concatenative Strategy", in *Proceedings of the North American Association For Computational Linguistics*, 2000, Seattle, United States.

Fassi Fehri, A., (1993), *Issues in the Structure of Arabic Clauses and Words*, Kluwer Academic Publishers, Dordrecht, Holland.

Mitamura, T., Nyberg, E.H., Carbonell, J. (1991), An Efficient Interlingua Translation System For Multilingual Document Production, in *Proceedings of the 3rd Machine Translation Summit*.

Nyberg, E.H., Mitamura, T. (1992), The KANT System: Fast, Accurate, High Quality Translation in Practical Domains, in *Proceedings of COLING'92*.

Schramm, G. (1962), An Outline of Classical Arabic Verb Structure, *Language vol. 38*, pp. 360-75.

Soudi, A., Eisele, A. (2004), Generating an Arabic Full-Form Lexicon for Bidirectional Morphology Lookup, to be published in *proceedings of Language Resources Evaluation Conference,* Lisbon, Portugal.

Soudi, A., Cavalli-sforza, V., Jamari, A. (2001), A Computational Lexeme-based Treatment of Arabic Morphology, in *Proceedings of The Arabic Processing Workshop, Association For Computational Linguistics*, Toulouse, France, 2001.

Soudi, A., Cavalli-sforza, V., Jamari, A. (2002b), A Prototype English-to-Arabic Interlingua-based MT system, in *Proceedings of the Processing of Arabic Workshop, Language Resources Evaluation Conference,* Las Palmas, Spain.

Soudi, A., Cavalli-sforza, V., Jamari, A. (2002a), The Arabic Noun System Generation, in *Proceedings of the International Conference on Arabic Processing*, Manouba University, Tunisia.

Soudi, A. (1999), Interfacing an Arabic Morphological Generator with an Interlingua-based Machine Translation System, ms. Carnegie Mellon University, USA.

Tomita, M., Nyberg, E.H. (1988), Generation Kit and Transformation Kit, Version 3.2, User's Manual, Technical Report, Carnegie Mellon-Center for Machine Translation.