Intuitive Coding of the Arabic Lexicon

Ali Farghaly

SYSTRAN Software, Inc
9333 Genesee Avenue
San Diego, CA 92121, USA .
farghaly@systransoft.com

Jean Senellart

SYSTRAN S A.
1 Rue du Cimetiere
95230 Soisy-sous-Montmorency, France
senellart@systran.fr

Abstract

SYSTRAN started the design and the development of Arabic, Farsi and Urdu to English machine translation systems in July 2002. This paper describes the methodology and implementation adopted for dictionary building and morphological analysis. SYSTRAN's IntuitiveCoding® technology (ICT) for facilitates the creation, update, and maintenance of Arabic, Farsi and Urdu lexical entries, is more modular and less costly. ICT for Arabic, Farsi, and Urdu requires the implementation of stem-based lexical entries, the authentic scripts for each language, a statistical Arabic stem-guesser, and separate declarative modules for internal and external morphology.

Keywords (machine translation, Arabic morphology, lexical entries, stem-based morphology, intuitive coding)

1 Intuitive Coding Technology

An effective way to reduce ambiguity and improve the efficiency of NLP systems is to incorporate domain-specific dictionaries (Farghaly & Hedin, 2003). In a general machine translation system, this involves the customization of the MT system to the particular corpora of a corporation. This customization process is typically performed by the system developers (Senellart et al , 2003).

Because domain specific information is propriety, the customization process is challenging. Corporate customers are reluctant to share such information with MT developers. Although it is important to note that most customers do not have the linguistic expertise needed to perform customization in-house. SYSTRAN developed the innovative IntuitiveCoding® technology (Senellart et al , 2003) to resolve this paradox.

Although SYSTRAN's current development of the Arabic system was done in-house, Arabic lexicographers with linguistic expertise were not readily available. We needed to make coding Arabic entries as intuitive as possible, in particular by starting with a stem-based Arabic lexicon, for increased productivity.

2 Stem-based Arabic Lexicon

The decision to start with Arabic stems rather than roots, eliminates the process of generating stems from roots. In other Arabic morphological analyzers (Beesley, 2001), the roots are entered manually as well as the morphological patterns. Such information is essential to generate Arabic stems and is complex to formalize. This decision is quite unique in the morphological descriptions of languages developed by SYSTRAN.

The main counterpart of this approach is an increased risk of typographical errors in the dictionary due to redundancy. SYSTRAN dealt with this matter by providing a "strict" coding frame for lexicographers with a "guesser" and validation features. At the same time, derivation is not directly described as such – verbal and derived nominals are distinct entries in the lexicon – though a link binds both. In a complete root-based system, a complex formalization is set up and encounters a large number of lexical exceptions (for example, inheritance of semantic features between stems). In SYSTRAN's system, the validation process looks for consistency of stems coded (for example, validating that the root is preserved in the different stems of a given verb). For a full discussion of the advantages of a stem-based dictionary over a root-based dictionary, see (Dichy and Farghaly, 2003).

Lexicographers were trained to enter stems, which are the words in the specific language. They do not need to consult grammar books to reach the underlying root and the patterns. Arabic native speakers struggle at school with these patterns known as الميزان الصرفي in Arabic grammar books. Another decision to eliminate the use of transliteration in the dictionary was made. As a result, lexicographers do not need to be trained in transliteration tables.

SYSTRAN uses an SQL database to maintain the dictionary, which automatically saves various versions for future translation quality comparisons and reinforces the consistency procedure on the database.

In our database an entry usually has six fields. The first three are for the lemma, part-of-speech and the type of the-part of-speech. Types represent sub-classification of major parts-of-speech. For example nouns have five types: common nouns, proper nouns, verbal nouns, present participle and past participle. There are two types of adjectives: base and comparative. There are several types of verbs, such as plain, aux_modal, aux_neg .etc. There is only one type of preposition. The last three fields are for the morphological, syntactic and semantic information. There is also a field for 'notes' in which the lexicographer may insert comments regarding the entry. Coding the linguistic information in the monolingual dictionary is a two- step process. The first step is completed when the first three fields (see Figure 1) are entered. Then, a morphological guesser is run to fill in the morphological field. The second step begins when lexicographers review the suggestions made by the guesser. If they agree with the suggested forms generated by the guesser, no additional work needs to be done. In the event they disagree, they make corrections and fill in the syntactic and semantic fields. In section 3, we show how the morphological guesser works.

قالَ	verb	plain
ليس	verb	Aux_neg
مُهَنِدس	noun	common
بَغْداد	noun	proper
مِن	prep	plain
قَليل	adj	base

Figure (1)
The First three fields of the dictionary

3 Statistical Arabic Stem-Guesser

Entering the morphological information proved to be very time consuming. For each entry, several stems have to be entered. This was done to avoid the use of morphological tables. The alternative is to enrich the lexicon. Figure (2) shows the morphological information of the verb زرع 'to plant'.

[perfect=زَرَعَ],[imperfect=يَزْرَع],[imperative=إزْرَع],[passperf=زُرِع],[passimperf=يُزْرَع]

Figure (2)
The Morpho field of "زرع"

The morphological field of nouns and adjectives contains forms for the singular masculine, singular feminine, plural masculine and plural feminine. In order to save time and reduce costs, a guesser was designed to automatically generate the different stems of each category. Only the rules that apply to the largest number of forms in a given category are used. Even though the lexicographers are aware that the guesser over-generalizes, they are 60% more productive. Figure (3) shows how the guesser over-generalizes and produces wrong forms that need to be corrected by the lexicographers.

[perfect=قال],[imperfect=يَقال],[imperative=إقال],[passperf=قال],[passimperf=يقال]

Figure (3)
The output of the guesser displays over-generalization

Generating forms with inflections for gender, number and person is performed by the internal morphology module presented in the following section.

4 Internal Morphology

SYSTRAN has two different modules for Arabic morphology: internal and external modules. The internal module generates all different inflectional patterns of a given stem. The input to the internal morphology module includes the stems in the morphological field. The rules are very simple. They go through the monolingual dictionary and retrieve the lemma, part-of-speech and the type. Next, the rules obtain the stem of the morphological field, identify the type of stem, and generate the correct inflected forms with tags that represent the morphological properties of the that form. Figure (4) shows the output of the inflected form كتبن 'they wrote'

كتبن verb plain كتب
+past+fem+3P+plural

Figure (4)
The output of the Internal Morphology Module

The internal morphology module in fact generates an inflected dictionary. As displayed in Figure (4), it generates the inflected forms exactly as they may appear in authentic Arabic texts. This module also provides the lemma, part-of-speech, type, gender, person and number tags since this information must be made available to the other modules in the MT system. This inflected dictionary is compiled using finite-state automata technology into a runtime dictionary.

The internal morphology module is thus simplified because many of the complex processes used to generate the verb stems were treated in the dictionary. For example, there is no need to design rules to generate the imperative form which vary from one verb to another because such forms are already accounted for in the dictionary.

However, rules for hollow and weak verbs, dual nouns forms, regular plurals, deletion, epenthesis etc. have been implemented in the internal morphology module.

5 External Morphology

It is very common in Arabic that words, or more accurately tokens, may exhibit the structure of a whole sentence. For example, the Arabic token شاهدناهم is translated into 'We saw them'. Therefore, this token must be decomposed into a verb, a subject and a direct object. Moreover, it is also possible that this token take a conjunction that will be attached as a prefix. The external morphology module is the component that decomposes a token into different parts-of-speech. The crucial difference between the internal morphology module and the external morphology is that the internal morphology works at the paradigmatic level; whereas the external morphology works at the syntagmatic_ level. Figure (5) illustrates the function of each.

	I	Internal	
External	هم	شاهدنا	و
	ني	يشاهد	
	ها	أشاهد	ف
	هن	سيشاهد	و

Figure (5)
The function of the Internal and External Morphology Modules

The different inflections that we see in the internal column represent variations of the verb with respect to tense, number and person. The relation of the members of the set under 'internal' is membership of the class of verbs; whereas the set across which the external morphology decomposes represent members belonging to different word classes. The first is a conjunction, the second is a verb and the third is a pronominal suffix. In the implementation, the external morphology module precedes the internal morphology because it feeds the lookup procedure. If decomposition is not done correctly, the lookup procedure will not match words that actually exist in the dictionary. There are cases where there will be multiple parses of a complex word. A complex word may decompose in more than one way. This is where disambiguation rules play an important role. External morphology rules written "intuitively" as combination patterns are described in Figure (6).

```
WAFA:=       <وَ.CONJ|فَ.CONJ>
KABILI:=      <كَ.PREP|بِ.PREP|لِ.PREP>
LI:=          <لِ.PREP>

# al+noun/det/adj/numeric
{WAFA}?_{AL}_<NOUN:-PROPERNOUN|ADJ
         |DET:QUANTIFIER|NUMERIC:CARDINAL>

# noun/adj-suffix
{WAFA}?_{NOUNADJ}_<PRON:PERSPOSS>
{WAFA}?_{KABILI}_{NOUNADJ}_<PRON:PERSPOSS>
```

Figure (6)
Sample of external morphology rules

Conclusion

SYSTRAN's Arabic-English machine translation system contains a dictionary of over 30,000 single-word entries. Terminology coverage of Arabic newspapers and Internet materials reaches over 90%. It currently provides adequate "gisting-level translation quality. We are developing a compound dictionary with a third level of morphology, a "compound morphology module". Analysis and synthesis rules are being added to improve the quality of the translation beyond the "gisting-level". Our approach has greatly accelerated the development of this system. Continued development on the lexical database, the syntactic module and quality assurance process are ongoing.

References

Beesley, Kenneth, 2001. "Finite-State Morphological Analysis and Generation of Arabic at Xerox Research:Status and Plans in 2001", ACL, Arabic NLP Workshop, Toulouse.

Dichy Joseph and, Farghaly, Ali, 2003. "Root & Pattern Vs. Stem: On what grounds should a multilingual database centered on Arabic should be built?" to be presented at IX Machine Translation Summit, New Orleans.

Farghaly, Ali and Bruce Hedin, 2003. "Domain Analysis and Representation", in . Handbook for Language Engineers, Ed. Ali Farghaly, CSLI Publications, Stanford, California.

Senellart, et al, 2003. "SYSTRAN Intuitive Coding Technology", to be presented at the IX MT Summit, New Orleans.