2011 International Conference on Biomedical Engineering and Technology
IPCBEE vol.11 (2011) © (2011) IACSIT Press, Singapore

MACHINE TRANSLATION FROM ENGLISH TO ARABIC

Mouiad Alawneh, Nazlia Omar and Tengku Mohd Sembok

Faculty of Information Science and Technology, National University of Malaysia, Bangi , 43600,Malaysia
National University of Malaysia (UKM), National University of Malaysia (UKM),
m_maradona86@yahoo.com

Abstract. Machine Translation has been defined as the process that utilizes computer software to translate text from one natural language to another. This definition involves accounting for the grammatical structure of each language and using rules, examples and grammars to transfer the grammatical structure of the source language (SL) into the target language (TL). This paper presents English to Arabic approach for translating well-structured English sentences into well-structured Arabic sentences, using a Grammar-based and example-translation techniques to handle the problems of ordering and agreement. The proposed methodology is flexible and scalable, the main advantages are: first, a hybrid-based approach combined advantages of rule-based (RBMT) with advantages example-based (EBMT), and second, it can be applied on some other languages with minor modifications. The OAK Parser is used to analyze the input English text to get the part of speech (POS) for each word in the text as a pre-translation process using the C# language, validation rules have been applied in both the database design and the programming code in order to ensure the integrity of data. A major design goal of this system is that it will be used as a stand-alone tool, and can be very well integrated with a general machine translation system for English sentences.

Keywords: MT, Agreement, Word reorder, Rule-Based, Example-based, Hybrid-based OAK, Parser, POS

1. Introduction

The current Machine Translation system facilitates the end user to understand the English textual sentences clearly by generating the precise corresponding Arabic language. Agreement is a basic property of language. In the most basic sense, agreement occurs when two elements in the appropriate configuration exhibit morphology consistent with their co-occurrence. Perhaps the most transparent case of this linguistic mechanism is number agreement between a subject and a verb: A singular noun in the subject position regularly co-occurs with a singular verb (e.g., "the dog runs"), and a plural subject noun regularly co-occurs with a plural verb (e.g., "the dogs run"). If the language has number marking on other elements, such as determiners or adjectives, these should also exhibit morphology that is consistent with their relationship to the subject head noun, and this co-occurrence relationship holds for gender and person agreement as well.

The modern Arabic dialects are well-known as having agreement asymmetries that are sensitive to word order effects. These asymmetries have been attributed to a variety of causes, first, by the analysis problems at the source language, second, the generation problems at the target languages. However, Arabic is not alone in showing word-order asymmetries for agreement, Similar asymmetries have been documented in Russian, Hindi, Slovene, French and Italian (Hutchins and Somers 1992). Languages are varied in the agreement requirements. Some of them like Arabic require number, gender, person, and case agreements while others need some of these agreements. Machine translation system develops by using four approaches depending on their difficulty and complexity. These approaches are: rule based, knowledge-based, corpus-based and hybrid MT, Rule-based machine translation approaches can be classified into the following categories: direct machine translation, interlingua machine translation and transfer based machine translation (Abu Shquier and Sembok, 2008).Our purpose of this paper is to design a hybrid-based (rule-based and example-based) framework based hence, to strike a balance between both approaches in the use of MT for the translation of

texts and to handle the problem of word agreement and ordering in the translation of sentences from English to Arabic.

2. Agreement and Word Reordering Problems in MT

In this section we will explore different areas that are expected to cause agreement and reordering problems during translation from English into Arabic. The test examples will be put to the Arabic MT system.

2.1. Adjective-Noun Agreement

This type of agreement is not found in English. Arabic however, requires that the adjectives agree in number gender, case and definiteness with nouns. but if the noun has one so must an attributive adjective **(**Mohammed and Sembok, 2007a). This is termed agreement in definiteness (Mohammad, 1990), and can be shown by the following examples:

1. house big (بيت كبير) [*bit kbeer*] a big house

2. the house the big (**البيت الكبير**) [*albit al kbeer*] 'the big house;

English adjectives are not marked for number or gender and so the predicative adjective does not agree with its subject. However, a predicate nominal must agree in number with the subject of its clause. Whereas Arabic adjectives require a number, gender and person agreements between the head word and the adjective. Here we will use the abbreviations sg, dl and pl to represent the singularity, dual and plural features respectively, and the gender features will be denoted as ***m*** for masculine and ***f*** for feminine.

Following are examples on adjective-Noun Agreement with THREE Arabic MT Systems:

- **A diligent rich handsome man**
  - (GOOGLE) وسيم رجل الاغنياء جدية [serious the rich(pl,m) man(sg,m) handsome(sg,m)]

- **A diligent rich handsome woman**
  - (SYSTRAN) إمرأة يهيأ غنيّة يجتهد [seeking rich good-looking woman]

- **Diligent rich handsome men.**
  - (GOOGLE) وسيم الاغنياء الرجال المتقن [the serious the men (pl,m) the rich(pl,m) handsome(sg,m)]

- **Diligent rich handsome women**
  - (SYSTRAN) نساء يهيأ غنية يجتهد [seeking rich good-looking women]

Examples Analysis:

None of the examples above have been translated accurately with Google or Systran as they did not make the adjectives agree in number and gender with their nouns, in example a with Google, the adjective rich that describe the noun man had been marked as a plural masculine adjective, where it should be singular as it describes the noun man which is singular, same case with example b, neither the adjective rich nor handsome had agreed in number and gender with the noun woman, in examples c, d and e, we can also notice the adjective-noun disagreement clearly, as for systran, it just translate awkward and ill-ordered translation in all of the examples above.

If the statement has more than one adjective that describes the same noun, then the same features of that noun will be used in the derivation of the all adjectives, for example:

The girl is strong and kind

Arabic translation is البنت شديده و لطيفه [*albnt shadeedah wa lteefah*]

Non-human nouns: If the noun that the adjective describes is plural and doesn't have the humanity feature then the singular female form of the adjective is used instead of the plural form.

Examples:

- The students are kind الطلاب لطيفون [*altolaab lteefoon*]
- The tigers are kind النمور لطيفه [*alasod lteefah*]

The second sentence uses the adjective لطيفه [*lteefah*] which is singular female form with a plural noun "*the tigers*", while in the first sentence the adjective used is in the plural male form لطيفون [*lateefon*] with "*the men*". The difference between the two sentences is the humanity feature in the first sentence, i.e., the men are human while in the second sentence the lions are not. This exception does not cover everything about the adjectives, but just a brief account to clarify the necessity for the agreement rules in MT.

2.2 Verbs-Subject Agreement

If a sentence contains a singleton subject noun phrase, how the verb is marked for agreement depends on the word order of the subject relative to the verb. In verb subject order the verb agrees with the subject only in gender and is marked in the singular, whether the subject is singular (1) or plural (2). Plural marking on the verb is only acceptable if the noun phrase is interpreted with contrastive focus as a SUBJ (3):

1. The boy wrote the homework كتب الولد الواجب
2. The boys wrote the homework كتب الاولاد الواجب
3. The boys wrote the homework (and not the girls) كتب الاولاد الواجب ولا البنات

In subject verb word order the verb agrees with the subject noun phrase in gender and number. If the subject is singular, the verb is marked as singular (4); if the subject is plural, the verb must be marked as plural (5); singular marking is unacceptable (6):

4. The boy wrote the homework الولد كتب الواجب
5. The boys wrote the homework الاولاد كتبوا الواجب
6. The boys wrote the homework الاولاد كتب الواجب

Arabic shows yet a more complex system in verb agreement than any other language, as the verb agrees with the subject in person, number, and gender. Both Arabic and English reflexives and possessives agree with their antecedents in gender, number (singular dual, or plural) and person. (e.g., He eats his food. or She eats her food.)

2.3 PRONOUNS

Only the pronouns he and she do not cause an agreement problem during translation into Arabic because they are clearly marked for number and gender. The other English pronouns you, they, it, I and we cause an agreement problem. This is due to the fact that the Arabic pronoun system differs from the English one in that the Arabic system includes a larger number of pronouns to allow for the distribution of features such as: singular, dual, plural, feminine, and masculine.

Test examples: Pronoun They with Tarjim:

a) They are two good boys. هم ولدان جيدون
b) They are two good girls. هم بنتان جيدات

Analysis:

The system uses the default masculine plural form of the pronoun in examples a and b, pronoun choice is wrong. The English pronoun it is not marked for gender. It is not clear whether it refers to a masculine or feminine object. Arabic, however, needs this distinction

3. Proposed Solution with Hybrid MT

Let us investigate the translation with Arabic MT system and see how it can handle the agreement and word-ordering, using hybrid – based MT following Methods steps:

STEP 1: Input the source text in English language

STEP 2: Pass the source text to the OAK Parser and get the output as (tagged POS)

STEP 3: From the output in 2, construct the English pattern in the format of the grammar table.

STEP 4: Check the procedure according to which EBMT is based is the following:

4.1 The alignment of texts.

4.2The matching of input sentences against phrase (examples) from stored database.

4.3The selection and extraction of equivalent target language or translated phrases.

4.4The adaptation and combination of translated phrases an acceptable output sentences.

4.5 When an example of the source language to be translated into the target language

Happens not to be found in the machine database go to step5.

STEP 5: Retrieve the record of this pattern from the grammar table in order to know the subject, verb, object, agreement requirements, and the equivalent pattern in Arabic language.

STEP 6: From the lexicon get the features and Arabic meaning for all words of the sentence.

STEP 7: Check for irregular word(s)

STEP 8: Apply the agreement rules for verbs and their subjects.

STEP 9: Apply the agreement rules for adjectives and the entities that they describe.

STEP 10: Apply modification rules on the object words.

STEP 11: Construct the Arabic text using the pattern exists in the grammar table.

STEP 12: Repeat steps 1 to 11on the next sentence.

4. CONCLUSION

Many shortcomings in the output of MT have been shown in this paper, due to either faulty analysis of the source language text or faulty generation of the target language text. Enhancement to the output can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. Fully automated, high quality machine translation (FAHQMT) has not yet been achieved. Yet there is a lot that we can do to improve the quality of MT output and increase its usefulness.

In this paper we have presented the necessity to handle both the agreement and the words reordering in the machine translation from English to Arabic. We proposed a hybrid-based approach to solve those problems; the paper has dealt with two features that greatly affect the output of MT, that are agreement and ordering problem which comes from the fact that different languages have different text orientation where some of them are left-to-right and others are right-to-left. The order of the words in the sentence is also different from one language to another.

5. References

[1] Satoshi, S. 2008, 'The manual of Apple Pie Parser v7.0' Computer science department, New York university.

[2] Attia, M. 2002. 'Implications of the Agreement Features in Machine Translation'. AL-AZHAR UNIVERSITY

[3] mohammd, and Sembok, T. 2007a. *'TOWARD FULLY AUTOMATED ARABIC MACHINE TRANSLATION SYSTEM*', IJCSNS International Journal of Computer Science and Network Security, 7 (5): 1-10.

[4] Franck, J. Lassi, G Frauenfelder, U. & Rizzi, L. 2006. 'Agreement and movement: A syntactic analysis of attraction'. Cognition, (101): 173-216.

[5] Hutchins, W. and Somers. L. 1992. ‘An Introduction to Machine Translation’. London: Academic Press. Love P.E.D and Irani Z. 2003. 'A project management quality cost information system for the construction industry'. Information and Management, 40(7): 649-661.

[6] Mohammad, M. 1990. ‘The problem of subject-verb agreement in Arabic: Towards a solution’, Amsterdam, Benjamins, Publishing Company: 95-125.

[7] mohammd, and Sembok, T. 2007b. ‘HANDLING AGREEMENT IN MACHINE TRANSLATION FROM ENGLISH TO ARABIC’. *The 1st International Conference on Digital Communications and Computer Applications (DCCA2007)*. JUST: 385 – 379.

[8] Trujillo, A. 1999. ‘Translation Engines Techniques for Machine Translation’*, Springer – Verlag Berlin Heidelberg*, New Work.