http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 4 (3): 256-261, 2005
ISSN 1812-5638

A Direct English-Arabic Machine Translation System

[1]Ahmad T. Al-Taani and [2]Eyad M. Hailat
[1]Department of Computer Sciences, Yarmouk University, Irbid, Jordan
[2]Al-Isra Private University, Amman, Jordan

Abstract: In this study, we present English to Arabic approach for translating a well-structured English sentence into a well-structured Arabic sentence. A dictionary is used to translate single words and a simple lexicon consists of only word categories and the meaning relative to category in appropriate format. The domain area is divided into two distinct areas: a set of English proverbs up to 184 proverbs from Al-Mawrid, English-Arabic dictionary and a set of 125 well-structured English sentences from many textbooks. Experimental results have shown that well-structured English sentences give better results that when translating proverbs.

Key words: Natural language processing, lexical analysis, parsing, Arabic language processing, artificial intelligence

INTRODUCTION

Machine translation includes any computer-based processes that transforms (or helps a user to transform) written text from one human language into another. Machine translation can be divided into three main trends: First, fully automated machine translation, which do the process of translation without intervention of human beings. Second, if the computer system does most of the translation but human may need for a help this is said to be Human-assisted machine translation. Finally, when the human does most of the work but uses one or more computer systems, mainly as resources such as dictionaries and spelling checkers, as assistants, is called Machine-aided translation[1].

Translation is a creative process that involves interpretation of the given text by the translator and translation would vary depending on the audience and the purpose for which it is meant; according to the context and situation. A problem of context can be solved by delimitating the subject domain so that machine works in a narrow subject area such as proverbs. Current research focuses on almost fully automatic systems, leading to extremely specific, task-dependent systems.

Several approaches for machine translation were used in several systems. Sakher Company is one of the first Arabic companies that work on this field. It developed software and a website that provides this facility.

Rached *et al.*[2,3] has proposed a system that translates Web pages from English to Arabic automatically. The system uses a commercial machine translation system to translate the texual part of a Web page. It then displays a Web page containing the Arabic translation with all tags inserted in the right places so that the layout and content of the original (English) page are preserved.

Al-Anzi *et al.*[4] has proposed a system to translate English Web pages to Arabic, a system has been developed at Kuwait University. The system partitions the English sentences into different parts according to where an HTML tag occurs. Then it translates the part of the English sentence independently of others and inserts the translation between the HTML tags that were present in the source. Experimental results showed that the system had faced difficulties when an HTML tag appeared inside a sentence.

Al-Mutarjim Al-Arabey[5] is another machine translation system. It is commercial software that translates English text into Arabic. This program is available from ATA software.

Beesley[6,7] described a finite-state morphological analyzer of written Arabic words. The system consists of the analyzer proper, running on a network server and Java applets that run on the user's machine and render words in standard Arabic orthography both for input and output.

The Speech Statistics (SpSt) project team[1] has designed a linguistic automaton aimed at natural language processing in a variety of forms. In addition to Arabic and German languages, the system handles text processing of a number of oriental languages.

Gey[8] states the requirements needed to implement machine translation. He suggested that one needs a bi-lingual dictionary of at least 250,000 words as

Corresponding Author: Ahmad T. Al-Taani, Department of Computer Sciences, Yarmouk University, Irbid, Jordan
Tel: +962 7 77438520 Fax: +962 2 7211128

the basic foundation, as well as general morphological software for both source and target language which will automatically parse sentences into their constituent adjective-noun-verb-object structure. Finally, one needs a transfer grammar, which maps the grammatical structure from the first language into the translated one.

In this study, a system for translating well-structured English sentences into Arabic sentences is presented. English sentences consist of lexical items. In traditional grammar, these are called parts of speech, or syntactic categories. The most important categories are the Noun (N), Verb (V), Adjective (A), Adverb (Adv) and Preposition (P). Lexical items combine to form larger units called phrases. These phrases could be nominal, verbal, adjectival etc. To identify the grammatical category of phrase we need to identify its head. When a phrase is headed by a noun, it is said to be a Noun Phrase (NP); when it is headed by a verb, it is a Verb Phrase (VP) and so on. These phrases combine to form larger units called clauses and sentences. In a sentence, some phrases (or even clauses) function as a unit, called a constituent. Sentences have an internal structure, that is, lexical items as well as the phrases that contain them are hierarchically organized. There are rules that regulate and govern the internal structure of phrases and sentences, called Phrases Structure Rules (PSRs). These PSRs are said to generate the sentences of the English language. PSRs are written using mathematical notation: XP→X. this equation reads, a phrase of type X must have X as its head. The arrow means consist of. Therefore if X=N, V, A etc. then XP=NP, VP, AP, respectively.

There are major kinds of phrases in English, which construct the well-structured English phrase. The major PSRs of English are:

- CP → C S.
- S → (NP/CP) Aux VP.
- VP → (AdvP) V ({NP/CP}) (PP) (AdvP)
- NP → (D) (AP) N (PP) (CP).
- PP → P (NP).
- AP → (AdvP) A.
- AdvP → (AdvP) Adv

The system is composed of two main phases. The first phase is the source language phase. This phase manipulates the English sentences generating the stem, the suitable grammatical category and finds the agreement for the phrase. The second phase is the target language phase. In this phase, we specify one Arabic meaning for each word and map the target language words according to the target language rules.

MATERIALS AND METHODS

The source language phase: This phase consists of four main steps:

Divide the input text into sentences and extract the sentence into words: The output of this step is a list that is ready to enter the system without any non-English characters. For more understanding we give the pseudo code for this step:

Step 0:
Step 0.1: Ask the user for input.
If the user chose an input in a text file, Then
The system will move to step 0.2.
If the user chooses a single sentence entered in the main form of the application
Then The system will goes to step 1.
Otherwise, The system will continue to next step.
Step 0.2: The input file will be broken down into sentences.
For each paragraph in the input text Do
0.2.1: Divide the paragraphs into sentences according to the punctuation marks and remove it from the output sentence.
Append the resulted sentence to the end of the list called FirstList
0.2.2: For each sentence in FirstList generated in step 0.2.1 Do
a. Extract the phrases according to English language Conjunctions mentioned above.
b. Translate the conjunction and remove it from the phrase.
c. Add the sentences resulted in the list called nonStemmedWordList.
d. Add the conjunctions into a list called ConjunctionList.

Generate the stem for each word: The stem of any word is the word with no additions. Additions may be suffixes, or prefixes. These are added to the stemmed English word to add additional meaning or to derive a word from another to complete the structure of a sentence. e.g. to derive a verb in past tense from a verb in present tense, we add the suffix -ed to the present verb and the resulted word tense would be past. In addition, the suffix -es or s

is added to words to derive others. We can add one of them to a verb to indicate that the subject of this verb is singular. There are common rules to derive words from others. Those can be found in the books that subject is vocabulary of a specific language, known that each language has its own different set of rules.

It is not necessary to find the stem for each word in the list. We may need not to carry out this overhead, because some words still have prefix and/or suffix and can be found in the dictionary. The input to this step is the list called nonStemmedWordList, which generated from the previous step. The outputs from this step are:

- A list called WordList, which contains the stem for each word in each sentence.
- A list called SuffixList, which contains the suffix that removed from words in the NonStemmedWorldList. If the word is a stem itself, then a special mark puts at the same position as the word in the NonStemmedWordList that is "_".

Step 1:
For each item x in the NonStemmedWordList Do
a. Split the sentences into its contained English words. The order of words is very important.
b. The number of words in the sentence will assign to the variable SentenceLength.
Step 1.1:
For each single word w_i in NonStemmedWordList Do
a. Search for the word w_i in the dictionary:
If the word w_i found Then Store the word in the list StmmedWordList at position i. Add the suffix in the list SuffixList at position w_i and Move to the next word (w_{i+1})
Else
If one of the suffixes located in the end of the word w_i
Then
Remove the suffix according to the groups of possible suffixes
If the word without any possible suffix Then
The word may be in other tenses, so search the simple past dictionary for the word,
If the word found Then
It is a simple past verb, so Extract the present word and store it in the position w_i in StemmedWordList,
Put ("-"+simple past verb) at the w_i-th position in the SuffixList.
Move to the next word (w_{i+1})
Else
search for the word in the past perfect field.
Do
If the word found Then It is a past perfect verb.
Extract the present word and store it in the position w_i in StemmedWordList,
Put ("=" + simple past verb) at the w_i-th position in the SuffixList.
Move to the next word (w_{i+1})

Find the most suitable grammatical category: This step is necessary because of the structure of the English language. In this step, the sentence is parsed into its constituent's noun, verb, adjective, adverb, etc structure that will be needed to apply PSRs mentioned above. The system works only for well-structured English sentences only, which satisfies the PSRs. We have used top-down parse tree technique to check the internal structure of the English sentence. The output of this step is the correct grammatical category of each word in the phrase and a top-down parse tree for the sentence structure.

The input to this step is the list that extracted in the previous step (StemmedWordList). The output of this step is a parse tree for the input string, on the other hand, tries are to avoid the ambiguity and specify the correct type of each word in the phrase, according to its position in the phrase. The final category according to the parse tree stored in the WordCatList.

Step 2:
For each Phrase in the list StemmedWordList from the previous step
Do
Step 2.1: For each word j in the phrase i in StemmedWordList
Do
If the word j is in the Main table, Then

Get the list of category(s) for the word j.
Store the list of types in a two-dimension array where the word j has a list of truly categories of the word at position j in the current phrase i.

Step 2.2: Now it is time to build the parse tree.
Initialize a variable called GPointer, the next word to be examined.
If the returned value of the functions EnglishSentence(GPointer) is true Then
The sentence is well structured and the exact category is determined.
Else
The sentence is not a well-structured sentence. Stop solving this phrase.

Find the agreement for the phrase: In this step, we do a morphological analysis and determine the person, number, gender and tense of the source sentence. These information will be needed when we want to output the target sentence in the target language, so we formulate the target one accurately and this will complete the meaning. The input to this step is the nonStemmedWordList and suffixlist.

Step 3:
Step 3.1: (Determining the PERSON)
For each word in the nonStemmedWordList Do
If the word is a pronoun Then Search for the pronoun and determine the person
If there are no pronouns in the phrase Then assign the 3rd person to the phrase and notify the user.
Step 3.2: (Determining the number)
If the word two appears in the phrase Then the phrase number is dual.
Stop.
Else If there is a pronoun in the phrase Then Determine the number according to the pronoun.
Else If the subject in the phrase has -s or -es suffix Then The number is plural
Else If the verb has a suffix -s or -es Then The number is singular
Step 3.3: (Determine the gender)
If a pronoun found in the phrase Then Determine the gender according to the pronoun
Else The search for the subject in the list of possible number proposed above.
Step 3.4: (Determining the tense)
If the verb has a suffix -s or -es Then the tense is present
Else If the verb has a suffix -d or -ed or the sign - Then the tense is simple past
Else if the prefix has the sign = Then the tense is a past participle

The target language phase

Specify one Arabic meaning for each word: In this step, the phrase is ready to be translated into the target language, word by word and in the same order as the source phrase. We search the database for the list of words that satisfy the query where the English word is the keyword with the exact category for this word. The output of this step is a list of Arabic words that gives the possible meanings for the corresponding English word.

We need only one Arabic word for each English word. Translators usually used to have the first Arabic meaning because many dictionaries put the more general and suitable meaning in order of occurrences in the language and the frequent of suitable meaning. For this reason, we employed simple rules, such as, the smallest word and the upper in the list, are preferable. The inputs to this step are the StemmedWordList and WordCatList and the outputs of this step are the list ArabicWordList and the list NonStructuredArabicList.

Step 4:
Step 4.1: For each word i at the position p in the StemmedWordList Do
Assert a query that the keyword is the word i, the category is that at position i in WordCategList,
Append the list returned from the database to the position i in ArabicWordList.
Step 4.2: Randomly, Pick one item from ArabicWordList at position i
If the item is a not single word then try again hundred times.
Insert the item resulted in the list NonStructuredArabicList.

Aligning the target words according to the target language rules: Now we have the raw material for a well-structured Arabic sentence, a set of lexical items not in the correct order. We have some rules in the Arabic language to align these words.

Arabic grammar books list the rules to construct the Arabic phrases. Those rules are written in Arabic language. For the source language phrase, the well-structured English sentence must consist of subject, verb and followed by an object. The input to this step is the list called NonStructuredArabicList that generated in the previous step.

The output of this step is the final sentence the system generates in the target language. In addition, the list StructuredArabicList contains the list for each phrase in the sentence.

Step 5:

The verb placed in the right most of the sentence.

If the subject and/or the object has the determinant THE Then

Add it to the front of the noun and the adjective.

Insert the noun phrase that act as the subject after the verb

Add the noun phrase that contains the object to the left most of the phrase.

EXPERIMENTAL EXAMPLE

Assume that the input phrase to the system is "The very smart boys ate the red apple"

Step 1: Divide the input text into words

Word	The	very	smart	boys	ate	the	red	apple
No.	0	1	2	3	4	5	6	7

Step 2: Generate the stem for each word

Word	The	very	smart	boys	ate	the	red	apple
Stem	The	very	smart	boy	eat	the	red	apple
Suffix	-	-	-	s	=ate	-	-	-

Step 3: Find the most suitable grammatical category

Word	The	very	smart	boys	ate	the	red	apple
Cate.	Det.	Adv	Adj	N	V	Det	Adj	N

Step 4: Find the agreement for the phrase

The output is: the person cannot be determined here; we assume the default, 3rd person.

The number is plural, the suffix -s added to the subject.

The gender is masculine, the word boy.

The tense is past; because of the "-" sign is the suffix.

Step 5: Specify one Arabic meaning for each word

For each English word with the specified category, the dictionary contains several Arabic meanings

Word	the	very	smart	boy	eat	the	red	apple
Cate.	Det.	Adv	Adj	N	V	Det	Adj	N
Arabic Meanings	ال - - - - -	جدا الى حد بعيد فعلا تماما - -	واخز،لاذع عنيف،قاس سريع،ناشط ذكي=ماهر=بارع وقح،أنيق ضخم،غال	غلام صبي ولد شخص خادم -	يأكل يلتهم يتأكل - - -	ال - - - - -	احمر=حمراء محمر خجلا=وردى متورد=محتقنه ضارب الى الحمرة إحمر متوهج	تفاحة شجرة التفاح - - - -

The result of this step gives the following possible Arabic translation for each English word in the sentence above.

English word	the	very	smart	boy	eat	the	red	apple
Arabic word	ال	تماما	الماكر	الشخص	يلتهم	ال	المتوهج	التفاحه

Step 6: Aligning the target words according to the target language rules

The result of this step gives the following possible Arabic translation for the English sentence.

English phrase	The very smart boys ate the red apple
Arabic phrase	يلتهم الشخص الماكر تماما التفاحة المتوهج

RESULTS

The domain area is divided into two distinct areas; the first one is a set of English proverbs up to 184 proverbs from Al-Mawrid, English-Arabic dictionary[9]. Al-Mawrid collects approximately 200 proverbs, that are used now and those forgotten, sorted in ascending alphabetic order. Some operations have done on this set of data by an English expert to be well-structured sentences. Second, a set of 125 well structured English sentences from many text books, including examples from English language grammar[10] and other books, to be input to our system.

A dictionary used to translate single words and a simple lexicon consists of only word categories and the meaning relative to category and in the appropriate format. A set of Arabic and English rules applied in several steps of the approach.

When evaluating the system, a percentage of approximately 57.3% of the sample proverbs was translated into Arabic and gave correct translation. The set of randomly selected English sentences gave 84.6% of the sentences were translated correctly.

DISCUSSION

Because of the following reasons proverbs did not give promising results, First, there is no specific structure for proverbs, because they are transferred from one generation to another generation orally. Second, ordinary people use proverbs extensively. Finally, proverbs are much related to the culture of some nations.

Machine translation is still one of the hottest topics in the computer field and in the computer market as well. The development of machine translation applications growth rapidly these days, because of many important applications depends on this application, one of them is Cross Language Information Retrieval (CLIR).

The proposed approach can be modified to include many computer applications as CLIR, knowledge-based systems and statistical machine translations. Moreover, the system can be improved to include non-well structured sentences. Future work to be done is the morphological step, Arabic words can be derived from other words, to give different meanings in different contexts.

ACKNOWLEDGEMENT

This study received financial support from Yarmouk University Research Council (Grant No. 2004/37), Irbid, Iran.

REFERENCES

1. Hutchins, W.J., 1988. Recent developments in Machine translation: A review of the last five years. In: Proceeding of the Conference on New Directions in Machine Translation, Budapest, 18-19 August, 1988.
2. Rached, N.Z. and A.G. Ahmed, 2001. An automatic English-Arabic HTML page translation system. J. Network and Computer Applications, 24: 333-357.
3. Al-Sikhan, A., R. Zantout and A. Guessoum, 1999. Automating the evaluation of machine translation systems lexicons: Arabic machine translation systems as case studies. In: Proceedings of the 7th International Conference on Artificial Intelligence Applications, Cairo, Egypt, February 3-6, 1999.
4. Al-Anzi, F., K. Al-Zame, M. Husain and H. Al-Mutairi, 1997. Automatic english/arabic HTML home page translation tool. In: Proceedings of the First Workshop on Technologies for Arabizing the Internet. King Saud University, Riyadh, Saudi Arabia, May 1997.
5. ATA Software, http://www.atasoft.com/products/mutarjim_v2.htm.
6. Beesley, K., 1997. Arabic morphological analysis on the internet. Xerox Research Center Europe, France. In: Proceedings of the International Conference on Multi-Lingual Computing (Arabic and English), Cambridge, UK, 17-18 April 1998.
7. Beesley, K.R., 1996. Arabic finite-state morphological analysis and generation. In: COLING-96 Proceedings, Copenhagen. Center for Sprogteknologi. The 16th International Conference on Computational Linguistics, 1: 89-94.
8. Gey, F., 2002. Prospects for Machine Translation of the Tamil language. Tamil Internet 2002, California, USA.
9. Baalbaki, M., 2003. Al-Mawrid, A Modern English-Arabic Dictionary: The Lamps of Experience, a Collection of English Proverbs with Origins and Arabic Equivalents. 37th Edn., Dar El-Ilm Lil-Malayen. Beirut, Lebanon, pp: 5-95.
10. Telfah, F.A., 1997. English for Academic, Political and Educational Purposes. 2nd Edn., Amman, Jordan.