ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Towards English to Arabic Machine Translation

Ehsan Abdulraheem Mohammed*
Department of Computer Science
Faculty of Information Technology, UKM
Bangi, 43600, Selangor, Malaysia
adimi.ehsan@yahoo.com

Mohd. Juzaiddin Ab Aziz
Department of Computer Science
Faculty of Information Technology, UKM
Bangi, 43600, Selangor, Malaysia
din@ftsm.ukm.my

Abstract**:** Nowadays, the understanding and generation of cognitive processes of natural language are becoming easier and better understood to perform machine translation. In this paper, we develop a system of machine translation to translate from English to Arabic, which runs on PC compatibles with English/Arabic interface. The system task was to analyze the natural language of English and Arabic to get accurate translation based on reordering per sentence at least one time. It based on identifying the Part Of Speech (POS) for each word in sentence from English dictionary, which used for reordering purpose. English dictionary also used to translate single word based on finding word meaning relative to categories (POS). The transfer of English words order in sentence from English structure to Arabic structure based on synchronize words between English and Arabic that based on matching between both language rules grammar. The meaning of words in sentences get from Bi-lingual dictionary that used to translate single word consists of only word meaning relative to categories (POS). This system applies on abstracts from European Psychiatry Journal domain that includes twenty (20) abstracts containing ninety five (95) sentences. The result obtained shows that the reordering rules is 81.9% accuracy on a translation from English Language to Arabic using abstracts from the European Journal of Psychiatry.

Keywords: MT; Natural language processing; Part Of Speech; Reordering

I. INTRODAUCTION

Machine translation that commonly known as MT in natural language. It is a sub-field of computational linguistic, which investigates the use of computer software to translate speech or text from one natural language to another. The efficient translation tools and fully automated in the current consensus should remain the primary goal.

The reason for build English- Arabic translation system is that Arabic is the lingua franca of the Middle-Eastern world. Presently, Arabic is national language in 21 countries with a combined population of 450 million with consider standard Arabic as their national language. Recently most of the researches in MT are mainly concentrated on the translation between English and Arabic that because English is a universal language and that will help in simplifying the Arab communication with other countries[1].

English and Arabic stem from different language families. Arabic alphabet consists of 29 characters, where the shape of each character depends on its position within a word [2]. Arabic is rich in morphological and syntactic structures [3-5]. In fact, Arabic as language has a complex morphology compared to English. Its also considered challenging in that it is a constructional or derivational language rather than a concatenative one [6].

The Jakob Elming (2008) concludes from experimentation, 91% of the sentences contain reordering in order to synchronize with an average of 2.96 reordering per sentence. As a result for this study, to get accurate translation must consider that the sentence needs more than four times as many reordering per sentence. Current paper focus on a problem to get high quality translation from English into Arabic of full text, that can be solved by first identify structure sentence format of English and match them with structure sentence format of Arabic. Then prepare translation rules set based on reordering structure format and that rules related to both of the synchronization of the structure languages.

A. English to Arabic reordering rules

Arabic-to-English direction constitutes most of the work in Arabic machine translation. English-to-Arabic direction is important to get the wealth of information of different domains, which is available, as well as the translation into Arabic that poses unique issues that are not present in the other direction. Badr et al. (2009) and Habash (2007) reported reordering rules for both Arabic-to-English and English-to-Arabic, by applying the rules, they select the ordering with the highest conditional probability.

It has been shown previously [7] that by starting from the top of the parse tree and descending recursively, the new order of every node and its children is determined by matching features of the node and children to the condition of an existing reordering rule. This process does not allow partial matching however, it is also deterministic in given the rules and adheres to the projectivity of the parse tree and utilizes a stepped back-off mechanism when a condition is not matched by attempts to match it with a different known condition. Firstly, it ignores the POS weight distinction and tries to match. If no matching happens, it drops the POS tag altogether and lastly ignores the syntactic label. In the case of complete mismatch, Arabic word order can be used through successful application of back-off which can bring the coverage to almost 100% to all rule sets examined. This back-off approach could be improved by allowing partial matches of the syntax-based rules used for re-ordering the English source to better match the syntax of the Arabic target [8]. English side of our corpora was reordered using predefined rules. These rules are based on Arabic syntactic facts. Reordering the English can be reliably done than other source languages such as Arabic, Chinese and German since the state of the English parsers are considerably better than parsers of other language in this research and complies with a report from Badr et al. (2009). We use the following rules for reordering at the sentence level and the noun phrase level was applied to the English structure:

a) NP: Reverse the order of nouns, adjectives and adverbs in the noun phase are inverted. A result of applying these rules, the phrase "very fast recovery" becomes "recovery fast very".

b) PP: All prepositional phrase of the form N1PrepN2PrepN3… Prep Nn is transformed to N1PrepN2PrepN3… Prep Nn unlessN1of N2of N3… of Nn is transformed to N1 N2N3… Nn. For instance, the phrase "geographical distribution of demand for psychiatric services" becomes "distribution geographical demand for services psychiatric". As well as if phrase containing "of" and before the prepositional "of" is noun phrase and these NP contains sequence of nouns and before this NP the definite article "the" then delete also definite article "the" with prepositional "of" like the phrase "The contribution of computerized mapping techniques" becomes "contribution techniques mapping computerized".

c) the: The definite article "the" is replicated before adjectives and this rule is applied NP in noun phrase after reverse. For instance, "The effective clinical treatment" becomes "the treatment the clinical the effective". The definite article "the"was also added after prepositional and before NP in genitive like the phrase "The new thrust in community care" becomes "The new the thrust in care the community".

d) VP: This rule transforms SVO sentences to VSO. All verbs were reordered on the condition that they have their own subject noun phase and are not in participle form since Arabic subject occurs before the verb participle. The following Example illustrates all these cases: "the patient need a special treatment" becomes "need the patient treatment special".

II. MATERIALS AND METHODS

A) Framework

The framework of the direct or transformer approach from English to Arabic MT system is given in Figure 1. Arrows indicate the flow of information. Oval blocks represent the basic modules of the system. Rectangle blocks represent the linguistic knowledge. This Figure also shows two main components: an analysis and a transfer component. Below list summarizes the translation processes of sentence in abstract based on Figure:

The translation process for each sentence in abstract based on above Figure 1:

- Pre-processor will read a sentence to return the words (lexical) of sentence to list for next process.
- Use English-Dictionary and small grammar to produce a source structure.
- Pass a package of rules, which transform the sentence into a target sentence, using where necessary information provided by the parsing process.
- The transformation rules are using the Bi-Lingual-Dictionary to find meaning for each word (lexical) in sentence in Arabic then align various rules, which is used reordering words.
- Target sentence (Arabic) will store into special database to retrieve later when the system is finished translating all sentences in abstract.

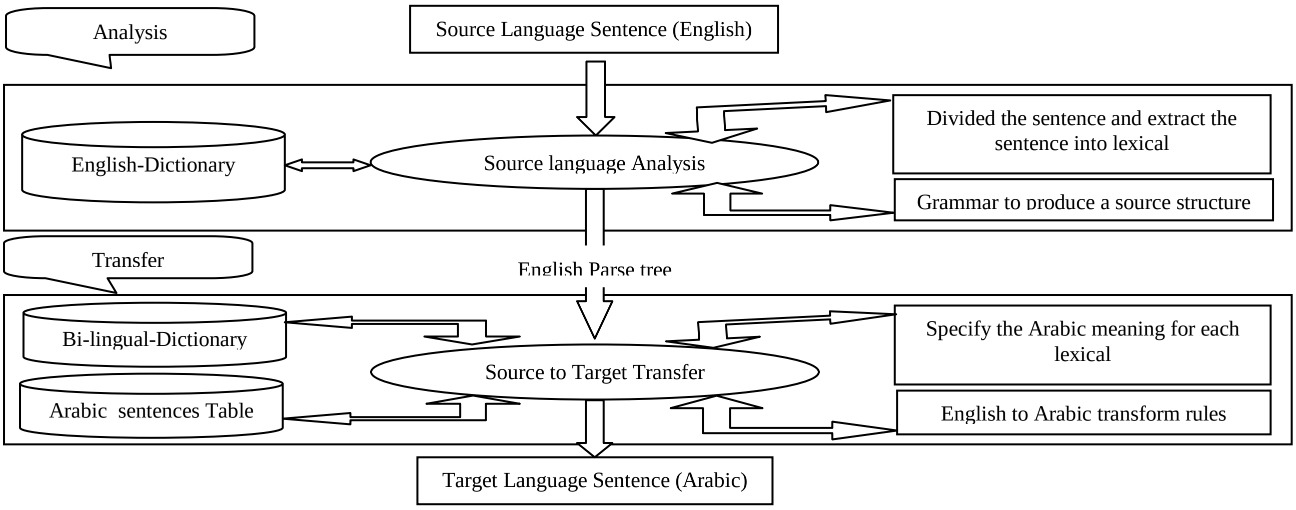
Analysis
Source Language Sentence (English)
English-Dictionary
Source language Analysis
Divided the sentence and extract the sentence into lexical
Grammar to produce a source structure
Transfer
English Parse tree
Bi-lingual-Dictionary
Arabic sentences Table
Source to Target Transfer
Specify the Arabic meaning for each lexical
English to Arabic transform rules
Target Language Sentence (Arabic)

Figure 1. Framework of English-Arabic Translation

A. Implement

The implement process is consist two phases that the source language phase and target language phase. In general, there are seven main steps. Figure 2 shows Block Representation of Reordering Algorithm, follow explains these steps. The system contains of two main phases. The first one is source language phase. In this phase divide English sentences into parts until reach to word level with manipulates that sentences (English) which generating in suitable grammatical category. The second phase is target language phase. This phase specifies one Arabic meaning for each word and aligns target language words according to the target language rules.

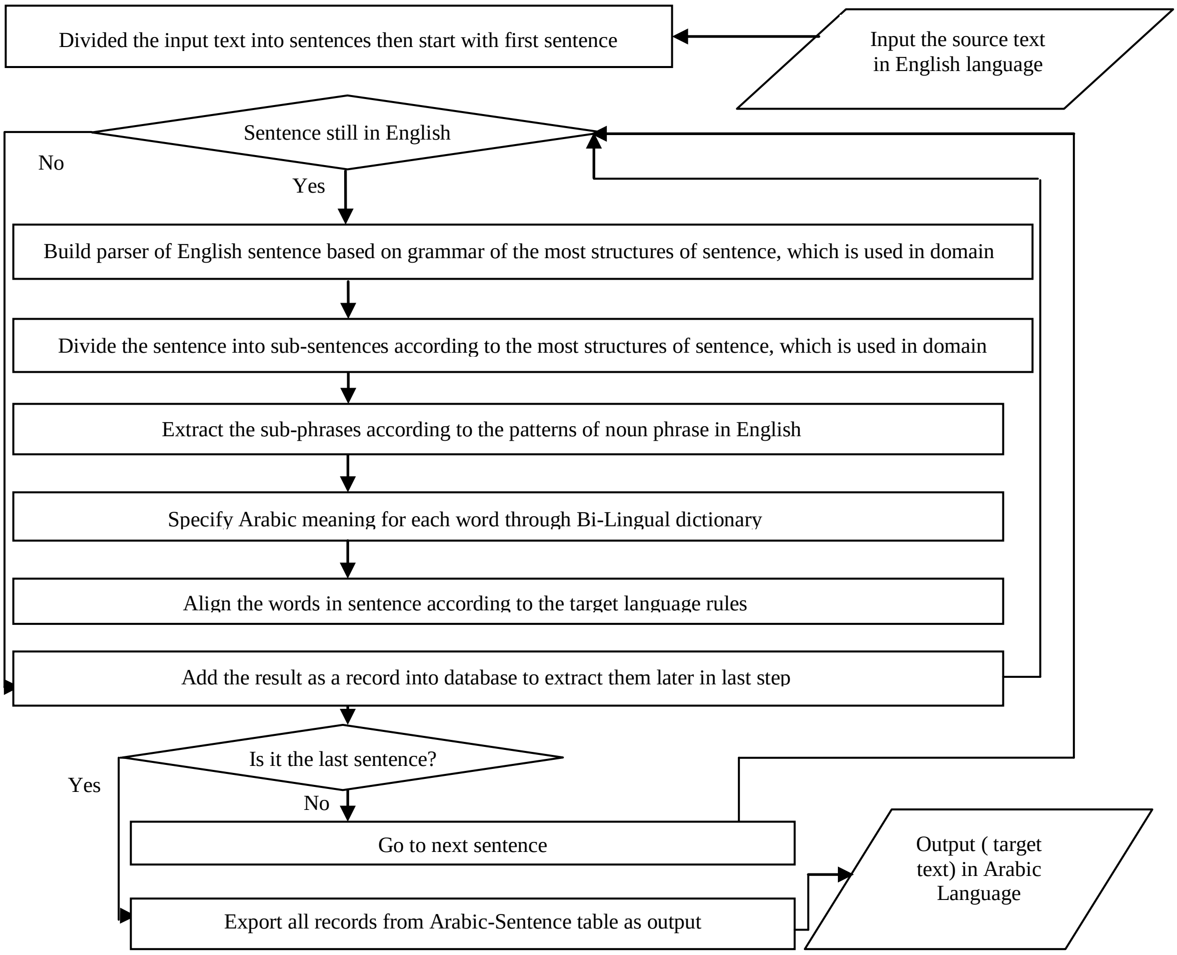
Divided the input text into sentences then start with first sentence
Input the source text in English language
Sentence still in English
No
Yes
Build parser of English sentence based on grammar of the most structures of sentence, which is used in domain
Divide the sentence into sub-sentences according to the most structures of sentence, which is used in domain
Extract the sub-phrases according to the patterns of noun phrase in English
Specify Arabic meaning for each word through Bi-Lingual dictionary
Align the words in sentence according to the target language rules
Add the result as a record into database to extract them later in last step
Is it the last sentence?
Yes
No
Go to next sentence
Output (target text) in Arabic Language
Export all records from Arabic-Sentence table as output

Figure 2. Block Representation of Reordering Algorithm

(1) *The source language phase:* This phase consists of four main steps:

Step1:

Divide the input text (abstract) into sentences. The output of this step is filling to each record in database with one by one sentence in order based on order sentence of source text (abstract), but those sentences in Arabic structure.

Step2:

Divide the sentence into sub-sentences. This step is important that for the structure of English. In this step, the sentence parsed into constituent's conjunction and punctuation structure, which will be need to apply English to Arabic reordering rules mentioned above. The system works only for structured English sentences only, which satisfied the English to Arabic reordering rules. We have used top-down parse tree technique to check the internal structure of the English sentence. The input to this step is the sentence that import from previous step (from matrix that store sentences). The output of this step is a variable that store all sub-sentences in sentence in order without any non-English character that means with target (Arabic) order structured.

Step3:

Divide the sub-sentence into phrases**.** In this step, the sub-sentence parsed into constituent's noun, verb, adjective, adverb, etc structure that will be need to apply English to Arabic reordering rules mentioned above. The output of this step is correct grammatical category of each phrase in the sub-sentence and top-down parse tree for each sub-sentence structure. The input to this step is the sub-sentence that extracted from previous step (from matrix that store sub-sentences). The output of this step is a variable that store all phrases in sub-sentence in order without any non-English character that means with target (Arabic) order structured.

Step4:

Divide the phrase into sub-phrases. In this step, the phrase parsed into constituent's noun, adjective, adverb structure that will be need to apply English to

Arabic reordering rules mentioned above of noun phrase. We also have some process before align these words in Arabic Language. For source language sub-phrase and other parts, the structure English sentence must add “the” in three situations. **a**. After preposition and before sub-phrase (noun phrase). **b**. before sequence nouns which prior by “of”. **c**. before sub-phrase (NP) when not prior by (the / a / an). It must remove “of” from phrases. The input to this step is the phrase that import from previous step (from matrix that store phrases). The output of this step is a variable that store all sub-phrases in order without any non-English character that means with target (Arabic) order structured.

(2) The target language phase: This phase consists of three main steps:

Step5:

Specify the Arabic meaning for each word. In this step, the phrase is ready to be translating into target language, word by word and in the same order in as the source phrase. We search the data based for the list of words that satisfy the query where the English word is the keyword with extract category (POS) for this word. The output of this step is a list of Arabic words that gives the possible meanings for the corresponding English word.

Step6:

Align the target words according to target language rules. Now we have raw material for a structure Arabic sentence, a set of lexical items not in the Arabic language correct order. We have some rules in the Arabic language to align these words. Arabic grammar rules list rules to contract the Arabic phrase. For the source language phrase, the structured English sentence must consists subject, verb and followed by an object, but the target language phrase, the structured Arabic sentence must consists verb, subject and followed by an object. The input to this step is the variable that generate in the previous step.

Step7:

Present the full abstract sentences in target language (Arabic). The output of this step is the final sentences which system generates in the target language structure. This output presents in interface after import them from database in order. Finally, delete all records from database that to free more space for next abstract process.

III. EVALUATION

The domain area is abstracts from European Psychiatry Journal include twenty (20) abstracts containing ninety five (95) sentences have been tested in order to verify the authenticity of computer translation and the result were compared with human translation. English dictionary is used to translate single word consist of only word categories (POS) and Bi-lingual dictionary is used to translate single word consist of only word meaning relative to categories in approach format. When evaluating the system, a percentage of approximately 81.86% of the sample abstracts was translate onto Arabic and give correct translation. The evaluation methodology will describe in the following steps[9-10] :

1) Run the system on the selected abstract text.

2) Compare the human translation with the system output.

3) Classify the problems that arise from the mismatches between the two translations.

4) Assign a suitable score for each problem that scores calls “Subscore”. A range of score between 0 and 10 determines the correctness of the translation. While 0 indicates absolutely incorrect translation and on other hand 10 indicates absolutely correct (matched) translation.

5) When a situation belongs to multiple problems compute its score average.

6) Determine the correctness of the test case by computing the percentage of the total scores first for each statement then for each abstract finally for all 20 abstracts

Table I and Figure 3 show the result of experiment for Systran, Google and Reordering Algorithm systems. We found the results of evaluation by calculate the average of abstracts (Total of Score (abstracts)) of system. The percentage of the total score for each system has been found by dividing the summation of total score *10 of each abstract by 20 (number of tested abstracts).

Figure 3. Score Percentage of Evaluation results

The results of experiment, the Reordering Algorithm system gets the highest score for each abstracts and that means this system is the best translation for abstracts in the domain rather than Systran and Google system. The highest translation score of abstracts is 87.5% and the lower translation score of abstracts is 76.7%. In general Google system is better than Systran, in some abstracts the systran translation score is little bit better the Google based on evaluation methodology and arising problems like abstract 4, 6, 10, 12, 13, 14 and 16. The highest translation score for abstracts is 81.5% and the lower translation score for abstracts is 73%. Systran system is the worse translation system rather than two previous systems.

The score of some translation of abstracts in this system is less than 50% like abstract 7, 17 and 18 that get 40%. The highest translation score for abstracts is 82.4% and the lowest translation score for abstracts is 40%.

Table I. Evaluation results

Abstract No.	**Systran Translation**		**Google Translation**		**Reordering Algorithm Translation**	
	Total Score	*Total Score *10*	*Total Score*	*Total Score *10*	*Total Score*	*Total Score *10*
Abstract1	7.43	74.3%	7.56	75.6%	8.40	84%
Abstract2	7	70%	7.67	76.7%	8.13	81.3%
Abstract3	7.15	71.5%	7.55	75.5%	8.16	81.6%
Abstract4	7.65	76.5%	7.33	73.3%	8.2	82%
Abstract5	7.56	76.5%	7.82	78.2%	8.08	80.8%
Abstract6	7.9	79%	7.56	75.6%	7.67	76.7%
Abstract7	4	40%	7.6	76%	8.38	83.8%
Abstract8	6.7	67%	7.3	73%	8.11	81.1%
Abstract9	7.56	75.6%	7.58	75.8%	8.13	81.3%
Abstract10	7.59	75.9%	7.44	74.4%	8.1	81%
Abstract11	7.35	73.5%	7.38	73.8%	8.2	82%
Abstract12	8.24	82.4%	8.15	81.5%	8.75	87.5%
Abstract13	7.52	75.2%	7.35	73.5%	8.4	84%
Abstract14	7.52	75.2%	7.38	73.8%	8	80%
Abstract15	6.09	60.9%	7.67	76.7%	8.11	81.1%
Abstract16	7.61	76.1%	7.55	75.5%	8.15	81.5%
Abstract17	4	40%	8	80%	8.1	81%
Abstract18	4	40%	7.9	79%	8.2	82%
Abstract19	6.14	61.4%	7.6	76%	8	80%
Abstract20	6.92	69.2%	7.5	75%	8.44	84.4%

IV. CONCLUSION

In this paper, we present our attempt to perform machine translation from English to Arabic. It has taken a most common structure sentence format in abstract of European Psychiatry Journal domain and that must considered for built system of MT from English to Arabic. In fact, until know fully automated, high quality machine translation (FAHQMT) has not achieved, but there is a lot that we can do to and increase its usefulness and improve the quality of MT output. This paper also show a system to solve the reordering problem that problem handle the words reordering in the machine translation from English to Arabic. The solution of reordering problem is focuses on the existing structure sentence format of our domain as well as identifying the Part Of Speech (POS) for single words and reordering the words in sentence for reorder purpose of English structure to Arabic structure and validates the reorder structure of whole sentence. We showed our system is promising and used to automate the translation of abstracts of European Journal of Psychiatry and that system get 81.86% of correct translation. To get 100% of correct translation that should need to solve the ambiguity of the meaning by create specialized lexicons, study the rest of structure format which used in our domain that because this paper focused on the most structure sentence format that used, but not all. These improvements will raise the 81.86% correctness of the translations.

V. REFERENCES

[1] Y. Salem, et al., "Toward Arabic to English Machine Translation," *ITB Journal*, pp. 20-30., 17 may, 2008. http://members.upc.ie/y.salon/pdf/Towards%20Arabic%20to%20English%20_ITB%20Journal-May-2008.pdf

[2] I. A. Jannoud, "Automatic Arabic Hand Written Text Recognition System," *American Journal of Applied Sciences*", vol. 4, pp. 857-864, 2007. doi:10.3844/.2007.857.864. http://www.scipub.org/fulltext/ajas/ajas411857-864.pdf

[3] K. Shaalan, et al., "Mapping Interlingua Representations to Feature Structures of Arabic

Sentences," *The challenge of Arabic for NLP/MT*, pp. 150-159, 23 October, 2006. http://www.mt-archive.info/BCS-2006-Shaalan.pdf

[4] K. Shaalan, et al., "Syntactic Generation of Arabic in Interlingua-based Machine Translation Framework" *in Third Workshop on Computational Approaches to Arabic Script-based Languages [at] MT Summit XII*, Ottawa, Ontario, Canada, 26 August, 2009. pp. 8. http://www.mt-archive.info/MTS-2009-Shaalan.pdf

[5] N. Adly and S. A. Ansary, "Evaluation of Arabic Machine Translation Based on the Universal Networking Language," presented at the Natural Language Processing and Information Systems, VerlagBerlin Heidelberg 2010. doi:10.1007/978-3-642-12550-8_20. http://www.bibalex.org/isis/UploadedFiles/Publications/NLDB-camera-final_1.pdf

[6] J. Elming and N. Habash, "Syntactic Reordering for English-Arabic Phrase-Based Machine Translation," *in Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Language*, Athens, Greece, 31 March, 2009, pp. 69–77. doi: 10.1.1.164.8046. http://www.mt-archive.info/EACL-2009-Elming.pdf

[7] N. Habash, "Syntactic Preprocessing for Statistical Machine Translation," *in Proc. of the MachineTranslation Summit (MT-Summit)*, 10-14 September, 2007. http://www.mt-archive.info/MTS-2007-Habash-1.pdf

[8] I. Badr, et al., "Syntactic Phrase Reordering for English-to-Arabic Statistical Machine Translation," *in Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens,Greece, 30 March – 3 April, 2009, pp. 86-93. doi: 10.1.1.1.154.4570. http://www.aclweb.org/anthology/E/E09/E09-1011.pdf

[9] A. A. E.-M. Mohamed, "Machine translation of noun phrases: from English to Arabic," master of scinece in computer engineering Master, Faculty of Engineering, Cairo University, Cairo, 2000.http://www.arc.sci.eg/NARIMS_upload/CLAESFILES/3805.pdf

[10] M. A. Shugier, "Word Agreement and Ordering in English-Arabic Machine Translation: A Rule-Based Approach.," The degree of doctor of philosophy Ph.D, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, selangor, 2009.