On One Model of English-Georgian Electronic Translator

MIKHEIL TUTBERIDZE
Department of Computer Sciences
Faculty of Exact and Natural Sciences
I. Javakhishvili Tbilisi State University
2, University st., 0143, Tbilisi
GEORGIA
mtutberidze@gmail.com http://tutberidze.gol.ge

Abstract: - In this work one model of English-Georgian Translator is considered. The algorithm of translation is constructed and the solutions of some incompatibilities are given.

Key-Words: - Translator, Algorithm

1 Introduction
The problem of translation is one of the known problems of informatics. One can find a lot of soft devoted to the solution of this problem. We can see the universal translators, which can translate from different languages to different languages. Interesting scientific works were created in this regard (see, for example [1], [2], [3]).

There are some methods elaborated for translating from one language to another. But still the problem of translation is not solved for many cases. For example, in case of Georgian language there is not any translator even from English language. This first of all is caused by the complexity of Georgian language and bad correspondence of grammar of these languages. We tried to find some regularity in English and Georgian languages, which would help in translation process. These regularities are mainly established in Grammar of these languages.

We tried to construct the model of the soft, which can perform translation from English to Georgian in case of sentences without special expressions. This model is based on the grammar of English and Georgian. We tried to find some correspondence between them and now we will consider this model.

2 Problem Formulation
First of all we must form the English-Georgian dictionary in the memory, which will be organized in the convenient way for searching. In the dictionary the physical connections between the words with common sense must be constructed.

Assume that we have an English sentence. We must translate it into Georgian. We have to find the corresponding words in Georgian and then try to connect them in the way not to lose the sense of the sentence. In this regard we will use the grammar instruments as well as the database of well-known combinations of words. Also we leave the opportunity to add some other analyzing method.

3 Problem Solution
3.1 Dictionary formation
English-Georgian dictionary will be constructed as array of addresses of structures, containing complex data. In particular, for each word it must be specified:
- The word itself (array of chars) and its Georgian equivalent (array of wide chars);
- Part of the speech to which the word belongs (char);
- Address of the structure, in which there is given the plural or singular form of the word (in case of noun, adjective);
- Address of the structure containing verbal forms of the word (in case of noun, adjective);
- Address of the structure containing verbal forms of the word (in case of noun, adjective);
- Address of the structure containing verbal forms of the word (in case of noun, adjective);
- Address of the structure containing verbal forms of the word (in case of noun, adjective);
- Address of the structure containing noun form of the word (in case of verb);
- Address of the structure containing noun form of the word (in case of verb);
- Address of the structure containing infinitive of the word (in case of verb);
- Address of the structure containing infinitive of the word (in case of verb);
- Address of the structure containing the adverbial form of the word (in case of adjective);
- Address of the structure containing the adjective form of the word (in case of adverb);
- Address of the structure, containing the infinitive of the word (in case of present, past and perfect participles).
- If the word has two or more meanings, for each meaning separate structure must be. In this case in each structure the reference to the next meaning must be contained.

The Georgian part of the dictionary is more complicated, because for each word its inflection and conjugation forms must be specified. In particular, for
nouns its inflections and plural (singular) forms must be specified. For adjectives and numerals its inflections in combination of noun must be specified. For pronouns there inflected equivalents must be specified. As for verbs, for them we have to specify their forms in twenty two different tenses, as well as its conjugations in six cases. Besides, we must specify the inflections for subjects and objects, connected with the verb for each twenty tenses. The formation of the structure array is performed from the binary file, which is also indexed. When a new word is added, the correspondent record is appended to the binary file, as for index, it is recreated. This takes some time, but on the other hand, the searching procedure is simplified.

3.2 Construction of the translation system
After retrieving the sentence from the text it is placed in the null-terminated array of chars. After this the number of words are established, the array of pointers is created and the addresses of words are saved. Then each word is compared with the dictionary for which the search process using the method of bisection is performed. For each word all cases of meaning are considered and the addresses of corresponding structures are retrieved.

After this we try to make a syntax research of the sentence. It is assumed, that in the sentence the syntax rules of word order are hold. First of all the type of the sentence must be explored whether it is confirmative, affirmative or declarative. For each type of the sentence the corresponding syntax order of words must be considered.

After establishing the type sentence must be checked whether it is compound or simple. For checking the compoundness of the sentence the pronouns or particles are searched for and if they are found in the proper context, then the type of compoundness is established and the sentence is separated into simple sentences. Otherwise the sentence is leaved as is.

The next step is exploration of the simple sentence. First of all the predicate must be found in the sentence. To do this, we have to make some steps:

If the words “will”, “shall”, “would”, “should” are met in the sentence it is assumed that tense of the sentence is one of Future or Future in the Past tenses. If the modal verb is met, it is assumed to be the predicate and its tense is considered to be continuous. If the word following the predicate is past perfect participle then the corresponding verb is considered to be the predicate in a passive voice and its tense is considered to be continuous. It must be also taken into account, that the participle in the last case can be preceded by adjective;

Otherwise if the verb “to have” (in any forms) is met in the sentence, it is assumed to be the part of the predicate and the words following “to have” verb are explored. If the word following “to have” verb is noun or pronoun, which may be preceded by attribute, it is assumed to be the second part of compound predicate. If the word following the predicate is past perfect participle then the corresponding verb is considered to be the predicate and its tense is considered to be perfect. It must be also taken into account, that the participle in the last case can be preceded by adjective;

Otherwise the verb not in infinitive or noun context is found and assumed it to be the predicate. Then the subject must be found. In case of confirmative sentence subject is considered to be before the predicate. If the sentence begins with the words “there” and “it”, then they are treated in the special way. Otherwise the noun or pronoun before the predicate is considered to be the subject. The adjectives or numerals before the subject are considered as attributes of the subject.

The noun after the predicate is considered to be object. If the noun is with preposition then it is considered to be indirect object, otherwise it is considered to be direct object.

As for the adverbs, if adverb is met in the sentence, it as assumed to be the modifier.

Such an approach is useful for us to translate into Georgian. First of all the correct tense must be chosen for the sentence in Georgian, which depends on the voice of the verb whether it is active or passive. Translation of passive voice into Georgian is more difficult than translation of active voice and therefore if some complications are met, the change of the voice will become necessary.

After the process, described above, the Georgian version of predicate is selected and put it in the correct tense. It is a difficult process, because in Georgian there are 22 tenses and one of them must be matched. Each of them has its own rules for using subject, direct object and indirect object. Besides, the right form of conjugation must be selected. The inflections of subjects and objects are chosen in correspondence with the tense of the verb. The attributes of the subject and object must also take inflection.

In English the sequence of words is strictly established. As for Georgian, there are no any restrictions. But for translation process some rules
must be established. The subject must be before the predicate, as for the objects, they will follow the predicate. The reflection of the objects determines whether object is direct or indirect. After this, if the initial sentence was compound and therefore separated into simple sentences, we have to merge their Georgian equivalents. The form of compoundness must be kept during the translation process. This is achieved by specifying the corresponding pronouns forming compound sentences.

3.3 Construction of the software

As one can see from the previous sections, the process of translation needs to perform a lot of laborious operations, which, of cause, needs a lot of time. Therefore, we must carefully select the platform, which will be used for creating soft. Use of the interpretive languages is ruled out. As for the database, it cannot be one of the existing databases and must be elaborated in accordance with the requirements with the soft.

We decided to write the soft using directly the library Windows API. This helps to achieve optimality. Probably we will have to load at least dictionary index file in the memory, if not dictionary itself. The soft can be realized as dynamic link library, which will make the soft easy to integrate within other soft. The soft will use the encoding UTF-8 and standard fonts such as sylfaen. The memory for the information, used by the system will be retrieved before the beginning of the translation process and probably will not increase during the process.

As for the database of the soft, we have to save it in tab delimited format in UTF8-encoding. This file will be indexed and this will help us when performing search procedure. No any requirement for the operating system parameters is needed.

4 Conclusion

The method described above serves a large interest in Georgian Specialists, mainly philologists. We make consultations with a number of specialists of English-Georgian Translation. We try to work with theoretical as well as practical specialists. By this time one group of scientists has created detailed English-Georgian Dictionary which we can use for our soft, but it must be converted into the electronic database which cannot be done without a big human resources. Nevertheless we work on the soft, which will realize all the said above. The soft is already working for some very simplified cases, making some translations, but there are many complications. Besides, we haven't the full database of words and word combinations, which is associated with large expenses. But we think that after finishing the soft we will have the powerful instrument, which will make sponsors to be interested in its development. We think that the main idea is already elaborated and has the chance to be realized in practice.

References: