An English to Turkish Machine Translation System Using Structural Mapping

Cigdem Keyder Turhan
Dept. of Computer Engineering
Middle East Technical University
Ankara 06531, Turkey
cigdem@ceng.metu.edu.tr

Abstract

This paper describes the design and implementation of an English-Turkish machine translation (MT) system developed as a part of the TU-Language project supported by a NATO Science for Stability Project grant. The system uses a structural transfer approach in translating the domain of IBM computer manuals. The general design of the translation system and a detailed description of the transfer component is presented in this paper.

1 Introduction

The TU-Language project sponsored by the NATO Science for Stability Programme was started in 1994 to establish computational foundations for the natural language processing research on the Turkish language with the collaboration of the Computer Engineering Department of Middle East Technical University, the Computer Science Department of Bilkent University and Halici Computing, Inc. The project attempts to perform extensive research on Turkish which will eventually lead to the development of an English to Turkish machine translation system, Turkish language tutorial system, a Turkish dictionary and other software tools to be used in further research.

In this paper, some issues in translating from English to Turkish languages, the translation domain, the outline of the machine translation system under development, and a detailed description of the transfer component will be presented.

2 Turkish Language

Morphology and syntax of Turkish are very different from English, therefore, the formalism used to represent English texts has to be altered significantly for Turkish text representation. The Turkish language is characterized as a head final language where the modifier/specifier always precedes the modified/specified. This characteristic also affects the word order of the sentences which can be described as SOV where the verb is positioned at the end.

Also, when compared to other languages, Turkish relies more on overt case markings which mark the role of the argument in a sentence. The case markings enables Turkish to have a relatively free word-order property where every variation in the word order in a sentence results in a different meaning.

In the MT system being developed, these and other different characteristics of the Turkish language are handled in the transfer and generation components.

3 Translation Domain

As more and more computer companies enter the Turkish market, a growing demand for English to Turkish translation of computer manuals has emerged. Other machine translation systems have also chosen the domain of computer manuals for their translation systems because of the relatively unambiguous and narrow sublanguage used (Tsutsumi, 1986). Also, in his research, Nasukawa (Nasukawa, 1993) concluded that the statistical analysis of the text in IBM computer manuals showed that 92.6 percent of the words in a computer manual are used in the same word sense which would significantly reduce the problem of lexical ambiguity resolution. Another advantage is that the material in a computer manual is observed to be written as clearly as possible in a relatively narrow area which will hopefully ease the difficult job of understanding and representing the input sentence.

As a result of these observations, the TU-Language project team has chosen the IBM computer manuals as their translation domain..

4 Machine Translation System

The English to Turkish MT system under development uses a structural transfer approach which has the following components. First, the English sentence retrieved from the IBM manual is analyzed

by the CLE parser (Alshawi and Moore, 1992) to generate an intermediate representation. This representation is mapped onto a recursively embedded case frame which is then input to the transfer module. The transfer module maps the input case frame into the target case frame which is then filtered to be transformed into the required input format of the target language generator. Lastly, the generator maps the Turkish case frame into the Turkish sentence which is then post-edited by a human translator to get an intelligible and accurate translation.

4.1 Analysis Phase

For analyzing the English input, the Core Language Engine developed by the SRI Cambridge Computer Science Research Centre was used (Alshawi and Moore, 1992). The CLE system has been trained to meet the lexical, syntactic and semantic demands of the IBM corpus. In CLE, explicit intermediate levels of linguistic representation are used in the different phases of the analysis. Following the syntactic and semantic analysis/synthesis which uses the unification-based approach, the quasi logical form (QLF) is developed. QLF can be described as a *contextually sensitive* logical form. Since the CLE system produces various parses for an input sentence, the best parse is filtered by the system which conveys the intended meaning of the sentence. Then the chosen representation is mapped into a case frame.

4.2 Transfer Phase

Experience with previous systems using the interlingua technique showed the significant complexity of extracting and representing *deep meaning* of a natural language text (Goodman and Nirenburg, 1991). Another major difficulty encountered with this approach is that the language specific attributes necessary to define the translation equivalents in the lexical and structural levels are neutralized in the interlingual representation thereby complicating the task of generation considerably.

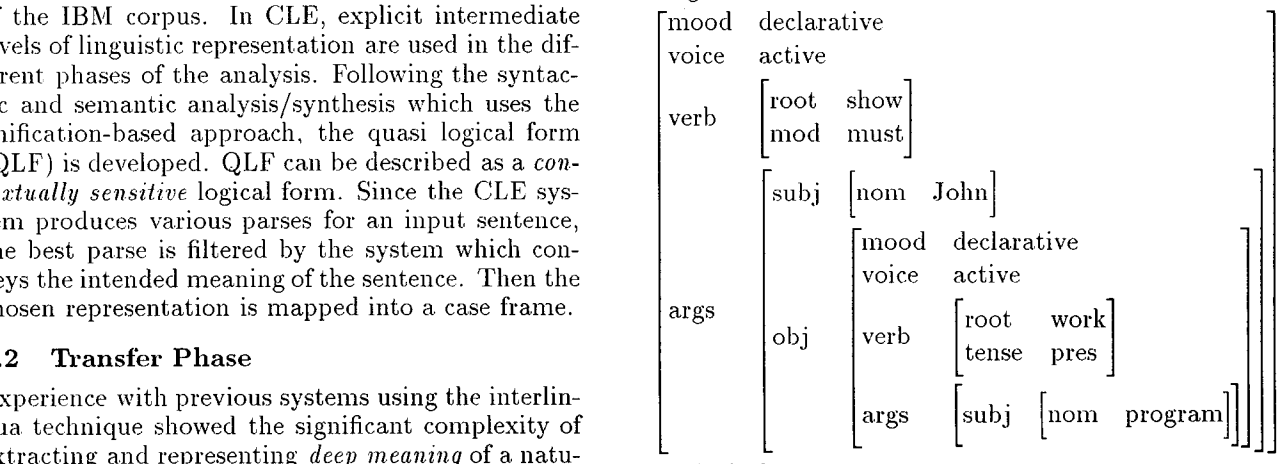
A similar problem occurred with systems using the transfer approach with deep semantic analysis such as the EUROTRA project (Johnson et al., 1985). Such systems were observed to be difficult to develop and maintain. To avoid these problems, the MT systems developed recently generally chose to use the straightforward transfer approach which relies on various types of lexical, syntactic information and a limited use of semantic analysis (Tsutsumi, 1986).

The system being developed as a part of the TU-Language project also chose the structural transfer approach with a minimal amount of semantic analysis. The transfer phase of our MT system performs structural transfer between the respective case frames of the analysed English sentence and targetted Turkish output. In a top-down manner, the transfer module transforms the English case frame or adds new information to the Turkish case frame in order to generate the equivalent Turkish noun phrase, clause or sentence with the aid of a transfer dictionary, and the transfer rules.

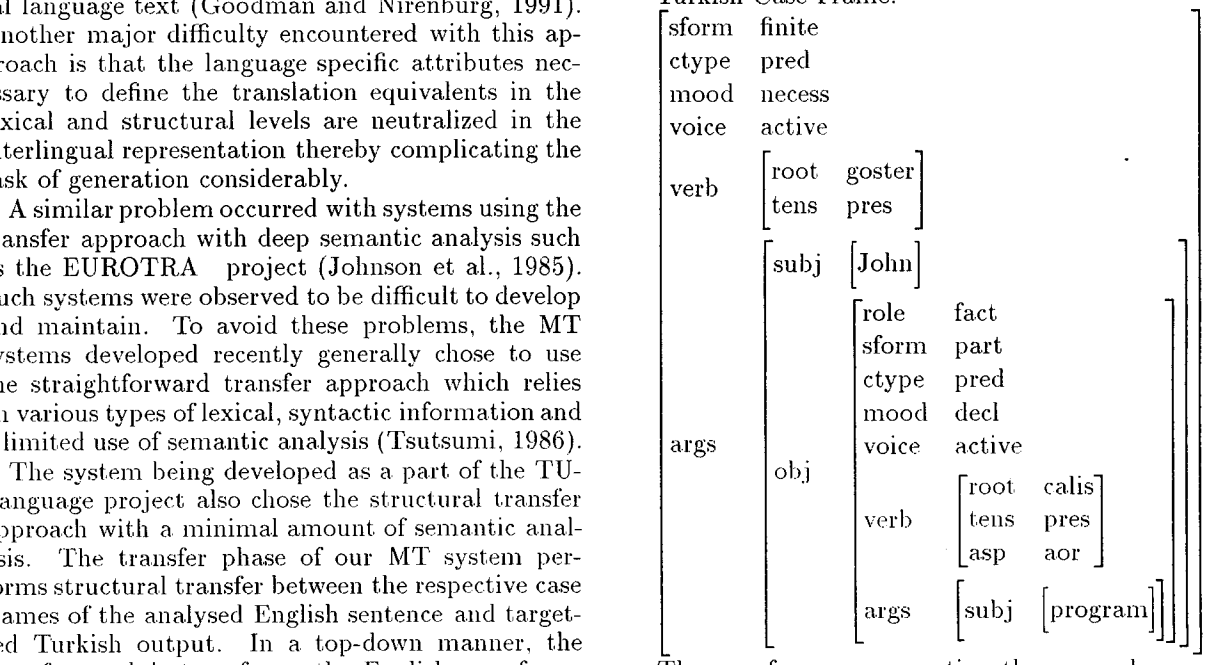
The English and Turkish case frames for clauses/sentences are generally similar to each other with differences seen in the sentence's mood and the verb's aspect and modality. Some information not extracted in the analysis phase such as the sentence form, clause type, role, etc. have to be determined in the beginning of the transfer phase and added to the Turkish case frame. An example sentence and parts of the corresponding English and Turkish case frames can be seen below:

(1) John must show the program works.
John goster+tns program+gen calis+tns
'John programin calistigini gostermeli'.

English Case Frame:



Turkish Case Frame:



The case frames representing the noun phrases of the English and Turkish sentences vary from each

other in a number of ways because the generator requires additional information to form an equivalent Turkish representation. For example, in the sentences below,

(2) That man writes programs.
O adam yaz+pres program
'O adam program yazar.'

(3) Programs were written for the project.
Program+pl yaz+pass+pst icin proje
'Proje icin programlar yazildi.'

Even though the word program is used in the plural form in both of the English sentences, the transfer module needs to determine the specificity of the noun phrase in question and send it to the generator which will accordingly output either the singular or plural form of the noun.

Some of the complex transfer issues presented by Lindop and Tsujii (Lindop and Tsujii, 1991) also arise in our machine translation system. These issues are handled with special transfer rules and transfer lexicon entries. In the beginning of the transfer phase, the exception rules are tested and eventually a checklist containing the problematic components of the input is generated. Some examples of these components are verbs which change meaning when used with different attributes, passive, existential or conditional sentences, relative clauses, idiomatic use of prepositional phrases, etc. As the transfer process continues, the checklist is referenced in order to block the default translation and handle the exceptions. The rest of the mapping proceeds in a straightforward fashion until all of the information in the source case frame is mapped onto the target case frame.

Some of the complex transfer issues handled in the transfer phase will be presented in this section.First, a significant amount of head-switching is performed to resolve the lexical and structural differences in the English and Turkish languages. In the example below,

(4) attempted execution
tesebbus calistirma
'calistirma tesebbusu : execution attempt'

execution is the head noun of the English phrase whereas *tesebbus* (attempt) becomes the head noun in the target phrase.

Another problem encountered in the transfer module is complex lexical transfer with category changes such as the example given below:

(5) John gave a weak cough.
John oksur+pst hafifce
'John hafifce oksurdu.:John coughed weakly.'

The adjective *weak* has to be mapped onto an adverb *hafifce* and the verb *give*'s default translation into the verb *ver* has to be blocked when it is used with the dependent noun cough. Consequently, the fitting target verb is found to be *oksurmek*.

Also, dependent on the verb, an object of an English sentence may be mapped to different case markings in Turkish.

(6) I hit the man.
vur+pst+pers adam+dat
'Adama vurdum'

(7) I shot the man.
vur+pst+pers adam+acc
'Adami vurdum'

As seen above,the object of the sentence *the man* is mapped either to an accusative marked object *adami* or a dative marked indirect object *adama* in the target sentence.

There are also some complex structural changes encountered during transfer. An English clause might be mapped into a Turkish gerund:

(8) While he was working
+ken calis+tns
'Calisirken'

Another example of a structural transformation encountered can be seen in active/passive forms of sentences. In the English passive form, the surface subject can correspond to both the direct object or the indirect object of the active form. Yet in Turkish, the surface subject of a passive sentence can only be the direct object of the active form. The difference between the two sentences is distinguished by the order of the phrases in the target sentence as seen in the example below:

(9) This program was given to the user.
Bu program ver+pas+pst kullanici+dat
'Bu program kullaniciya verildi.'

(10)
The user was given the program.
Kullanici+dat ver+pas+pst program
'Kullaniciya program verildi'

In both of the Turkish translations, the surface subject is *program* whereas the surface subject changes in the English inputs.

The order of the words in the output sentences are determined by the topic and focus features of the target case frame which are mapped during the transfer phase. In the first sentence, the topic is found to be *program* and the focus is *kullanici*, whereas in the second sentence the topic and the focus are *kullanici* and *program*, respectively.

The transfer module also attacks problems related to sentential transformation such as the ones required in the example below:

(11)
There are programs in the disk.
var program+pl disk+loc
'Diskte programlar var'

Parts of the case frames for the sentences above are as follows:

English Case Frame:

$$\begin{bmatrix} \text{mood} & \text{decl} & & \\ \text{voice} & \text{active} & & \\ \text{verb} & \begin{bmatrix} \text{root} & \text{be} \\ \text{tens} & \text{pres} \end{bmatrix} & & \\ \text{args} & \begin{bmatrix} \text{subj} & \begin{bmatrix} \text{nom} & \text{entity} \\ \text{det} & \text{there} \end{bmatrix} \\ \text{arg2} & \begin{bmatrix} \text{nom} & \text{program} \end{bmatrix} \end{bmatrix} & & \\ \text{adjs} & \begin{bmatrix} \text{place} & \begin{bmatrix} \text{nom} & \text{disk} \end{bmatrix} \end{bmatrix} & & \end{bmatrix}$$

Turkish Case Frame:

$$\begin{bmatrix} \text{sform} & \text{finite} & \\ \text{ctype} & \text{exist} & \\ \text{mood} & \text{decl} & \\ \text{voice} & \text{active} & \\ \text{verb} & \begin{bmatrix} \text{root} & \text{to-be} \\ \text{tens} & \text{pres} \\ \text{asp} & \text{aor} \end{bmatrix} & \\ \text{args} & \begin{bmatrix} \text{subj} & \begin{bmatrix} \text{program} \end{bmatrix} \end{bmatrix} & \\ \text{adjs} & \begin{bmatrix} \text{place} & \begin{bmatrix} \text{disk} \end{bmatrix} \end{bmatrix} & \end{bmatrix}$$

Other problems encountered in the transfer phase are the lexical gaps, idiomatic uses of phrases, and lexical disambiguation by syntactic or semantic content.

With all the complex transfer issues resolved in the transfer phase, the corresponding Turkish case frame is generated which is then translated from its Prolog notation into the Lisp notation required by the generation module.

4.3 Generation Module

The generation component of the system is based on the GenKit environment developed at the Carnegie Mellon University - Center for Machine Translation which provides facilities for a unification-based generation grammar environment (Hakkani et al., 1996).

As input, the generator receives a recursively embedded target case frame representation where all the lexical choices have been made, and produces the Turkish sentence conveying the same meaning.

Since Turkish has complex agglutinative word forms, a separate morphological generator handles the proper morpheme selection, vowel harmony, etc. to produce the surface form of the generated words.

The Turkish sentence output by the generator is post-edited by a human translator to ensure accuracy and intelligibility of the target sentence.

5 Conclusion

In this paper, an English to Turkish MT system using the structural transfer approach with a limited amount of semantic analysis has been described.

The structural transfer method which uses the recursively embedded case frames as intermediate representation proved to be very suitable in the application of English to Turkish machine translation. The greatest difficulty encountered with this approach is handling the complex transfer issues that arise due to the differences between the two languages.

Hopefully, the introduction of the English to Turkish MT software into the Turkish market will meet the growing demands for accurate, fast and high-quality translations in the field of computer manuals. Depending on the success of the system, the lexicons and the transfer module might be modified to tackle other translation domains in the future.

6 Acknowledgements

Helpful comments of Asst.Prof Cem Bozsahin and Assoc.Prof.Mehmet R.Tolun are gratefully acknowledged. This work has been supported by the NATO Science for Stability Project TU-LANGUAGE.

References

Hiyan Alshawi and Robert C. Moore. 1992. Introduction to the CLE. In Hiyan Alshawi, editor, *The Core Language Engine*. The MIT Press, Cambridge, Massachusetts.

Kenneth Goodman and Sergei Nirenburg ed. 1991. *The KBMT Project: A Case Study in Knowledge-Based Machine Translation*. Morgan Kaufmann, San Mateo, California.

Dilek Z. Hakkani, Kemal Oflazer and Ilyas Cicekli. 1996. Tactical Generation in a Free Constituent Order Language. In *Proceedings of 8th International Workshop on Natural Language Generation*, Sussex, UK, June.

Rod Johnson, Maghi King and Lois Tombe. 1985. EUROTRA: A Multilingual System Under Development. In *Computational Linguistics*, 11:155–169.

Jeremy Lindop and Jun-ichi Tsujii. 1991. *Complex Transfer in MT: A Survey of Examples*. CCL/UMIST Report No:91/5.

Tetsuya Nasukawa. 1993. Discourse Constraint in Computer Manuals. In *Proceedings of the TMI 1993*, pages 183–193.

Taijiro Tsutsumi. 1986. A Prototype English-Japanese Machine Translation System for Translating IBM Computer Manuals. In *Proceedings of Coling 1986*, pages 646–648.