Georgetown University
Institute of Languages & Linguistics
Research Project in
Machine Translation

Seminar Work Paper:
MT - 35

STATEMENT OF THE PROPOSED METHOD FOR MECHANICAL TRANSLATION

by

Ariadne W. Lukjanow

Working on the problem of Mechanical Translation (hereafter called MT) and considering the relationships and differences between the Russian language as a source language and English as a target language, I came to the conclusion that creative human analysis is an unavoidable part of translation. Since no creative work can be expected from the machine, all the preparatory work has to be performed beforehand by human beings and the tasks for the machine reduced, if possible, to matching of words and code digests, adding and subtracting of parts of the code and rearranging code units.

Consequently, the machine has to receive the following:

a. detailed Russian and English vocabularies.
b. code.
c. rules for operation.

These three requirements are true for any system or method of MT. The question is only what kind of analysis has been used, how the vocabularies have been compiled, what system of code has been devised and for that purpose, and how complicated or simple the rules for operation are.

To be the simplest method for achieving MT is along the lines outlined below.

The Russian language as a source language is the given data for the purposes of MT. It cannot be pre-edited or changed; its rules, regularities and irregularities are expressed in grammar. The evaluation of the specifications of the source language determines the procedure of research. In my case I have decided to use the difference between the two languages, the "ambiguities" of the source language, and make them solve the problem of Mechanical Translation.

One of the most striking differences between the two languages is the fact that Russian is an inflected language and English is not. After careful consideration of this factor, I came to the conclusion that I could make this "handicap" (inflection and ambiguity) work for me instead of considering it as a difficulty. The inflection in the language creates a very clear and easily detectable relation between the words in the sentence; it binds words together like the links of a chain within the sentence. The form of one word determines the form of the next. If this relationship can be utilized and expressed in a code, the words will select each other automatically. It is like a dial phone: every constituent of the sentence demands and thus dials certain specifications in the next until the unit is complete - by dialing the correct unit we are getting the party on the other end - the corresponding words of the target language. For example, any combination of preposition-adjective (several adjectives)-noun (several nouns), agree in case if they "belong together". Thus, the case of the preposition selects the case of the following adjective (or noun); the adjective selects the case of the following noun: noun-noun combinations agree in case if separated by a comma and disagree in a certain manner if they are not separated by a comma; verbs can demand a certain case from the following preposition, adjective or noun and so on. This dependence of one item of the sentence on the next item is always present within a Russian sentence or phrase. This can be illustrated by a simple sentence:

Я говорю о работе

Я - pers. pronoun, 1st pers. nom. = I

ГОВОРЮ - verb, 1st pers. present tense, no demands for the case of the next word = speak

O - preposition, demanding acc. or locative case from the following word and translated by two English prepositions, one for the locative case, and one for the acc. = about and against.

работе - sing, noun in dative or locative = work (locative), to work
(dat)

Together these items give us the following:

about (locative) work (locative)
I speak
against (acc.) to work (dative)

The case markers of the preposition and noun are matched and thus select each other.

Result:

I speak about work.

Now the sentence "I speak against work" could be possible too, but not in Russian, not if the same constituents are used. If we select the acc. case for the word work **(работу)** **we** would have to use the preposition ПРОТИВ instead of о The preposition ПРОТИВ = against, however, demands the genitive case from the following noun. Also, if we wish to use the proper Russian we would have to replace the verb with another verb of similar meaning: otherwise the sentence would be incorrect. At the same time, the preposition О with a different noun would mean against and demand the accusative case from the noun. Example:

Я опираюсь о стол

Я - pers. pronoun, 1st pers., nom. = I

ОПИРАЮСЬ - verb, 1st pers., pres. tense ; transitory verb and as such in this form in the absence of negation requires the acc. case from the following lexical item. = lean

О - prep., demands acc. or locative cases = about and against.

стол - noun, sing., acc. or nom. cases = table.

Together:

about (locative) table (nom.)
I lean (acc.)
against(acc.) table (acc.)

Result of matching:

I lean against (the, a) table.

Consequence: because of a certain interdependence of lexical items within a sentence or phrase in the source language only one translation of each item (with correct meaning and form) is possible despite the fact that even in these simple short sentences at least two choices were necessary.

As a next step I shall give an example as to what I mean by subtracting or omission of code units (and with it omission of lexical items in the target language). We shall take an ambiguous form of the Russian noun дома - nom. and acc. plural, genitive sing. = houses and of house (note: all prepositional cases for nouns are coded as preposition plus noun) and the preposition для, which demands the genitive case and translates as for. After matching we shall have for of house. According to the rule for omission or subtracting which states that if one preposition fellows another and the second one is part of the next lexical item (result of declension or case) it should be omitted.

Now, why is it necessary to have these case prepositions in the code and target language glossary and why not list nouns or adjectives without prepositions? The answer is as follows: in Russian there are two cases (for all cases but nominative and locative) prepositional and non-prepositional. In case of the absence of the preposition in Russian, the case preposition would be absolutely necessary for the translation and will be present in the code as well as in the glossary, thus eliminating insertions or decisions as to whether or to not to translate a Russian suffix or how to translate it. By translating Russian words in their multiple forms, I translate the suffix, affix and stem at once and have the whole listed in the code and glossary. Then, by matching I can keep them or discard them. Presence of the preposition as an independent lexical item automatically discards the following preposition if the latter is part of the next lexical item (noun or adjective).

Adding. By using the term adding I do not mean that something is added to the code digits which constitute the sentence. It is, for instance, when one lexical item (code unit)plus another lexical item(code unit) equals a third lexical item. This is necessary for some lexical items: two Russian adverbs standing together in the sentence have to be translated together or they lose their meaning: chemical terms in Russian, when used in groups and designating chemical substances, have to be translated as a unit (oxide ferrum is ferric oxide in English). However, if the terminology of chemical elements is marked in a special way and their "belonging together" is easily established, by the matching of case-markers, they can be translated as a unit (item A plus item B equals item C, and C is what is found in the glossary.

Since this method is based on the interrelation of items in the sentence, which in turn is expressed by standard Russian grammar, I used the data and terminology of standard Russian in my code. As I see it, my code is not an intermediate language in the strict sense; it is rather a device to achieve selection, arrangement and replacement through matching and thus translation. I used English as well as Russian grammar, which means that each language had to be analysed in terms of the other, with the result that both grammars were, as in mathematics, reduced to a "common denominator". For example, the target language received the equivalents of two additional cases, the instrumental and locative, so that certain prepositions plus nouns or adjectives were treated as adjectives and nouns in these cases: some source language adjectives became nouns etc. Then, the inflection of Russian serves for selection, while the rules derived from English grammar serve for omission and arrangement. Using the terminology and rules of standard Russian grammar, I had to devise a code which would bring forward, signalize and mark the following:

a. Parts of speech, case, tense, sing. or pl., gender, 1st, 2nd, 3rd pers. etc, and their possible demands to the following item.

b. Numerical digits, in order to be able to distinguish one lexical item from another.

c. Markers for relation to the subject translated in case of multiple translation of one single item or to be able to pair lexical items.

<u>Preliminary and partial</u> (arbitrary) <u>code symbols:</u>

SI	Noun
A3	Adjective
L2	Preposition
N4	Adverb
V5	Verb
T5	Auxiliary verb
P6	Participle
D7	Past participle
M8	Pronoun
C9	Conjunction
0	Negation

i	nom.
g	gen.
d	dat.
v	acc.
t	instr.
r	locative

m	masc.
f	fem.

c - neutr.

p - pl.

-2 - pl. for verbs

y - selection for translation
x - chemical terms
sp - special cases (in idiomatic prepositional combinations of certain prepositions plus certain nouns that serve as adverbs and have to be translated as such).

Further signalization should be introduced in the code to achieve a wider and better selection of prepositions. If nouns (only certain ones) have signals specifying their relation to time and space, the selection of translation of prepositions can be made from a larger number of possible translations by matching not only case-markers, but also tense additional signals.

Every item in the source vocabulary has been put on a card, coded in accordance with what I have said above. Units of this cod: have been put on cards with corresponding equivalents in the target language. In this manner the source language vocabulary is connected to the target language counterpart by means of the code.

SUMMARY

I. The proposed method is based on:

a. Analysis of both source and target languages in terms of each other in order to achieve the proper translation and correspondence of grammatical forms and meaning of lexical items entered in the glossaries.

b. The combined grammars of both source and target languages. **The** inflection in the source language gives a very clear and easily detectable relationship between lexical items in the sentence, the form and meaning of one lexical item determines the grammatical form and meaning of the other

Thus by expressing the inflection of the source language in symbols of the code and interdependence of the lexical items in the sentence in operational rules, automatic selection of lexical items is achieved through matching code symbols. The grammar of the target language is utilized in the same code and serves as a basis for operational rules to facilitate the necessary arrangements, omissions and replacements of lexical items.

2. A given item in the source lexicon, with its affixes of declension, conjugation, or with multiple entries when multivalent, is entered in the source lexicon with a systematic code based on grammatical categorizations (in the case of declensions, conjugations and other forms of affixation) and meaning distinctions in the case of multivalent items.

3. In cases of declensional or conjugational forms or multivalence, a series of code units (digits and letters) appears after a given lexical item.

4. The same code is entered with the corresponding target lexical items. In this way every lexical item of the source language is bound to the corresponding (grammatical form and meaning) lexical item of the target language by means of the code.

5. In cases where a source lexical item carries several code groups (units) because of ambiguity, a selection of the appropriate code group is effected by a matching procedure based on the presence or absence of corresponding code diacritics after the word or words following. This results in proper selection and thus removes ambiguity.

6. The translation of a source item will be effected by matching of the codes entered with the source lexicon and with the target lexicon, thus permitting the elicitation of the appropriate target item.

7. Another feature of the technique is the temporary storage of the code items. When all items of the given text have been stored the required selections and manipulations can be performed by application of operational rules.

Then the group of code items produces as output in the temporary memory should by matched to the target language glossary and the result of this last matching will produce the translation.

8. The sequence of operations:

a. Match the lexical items of the text to the lexical items of the source language glossary; read all code groups listed with each lexical item; store code groups in the temporary memory.

b. Read code groups stored in the temporary memory from left to right, matching, selecting, discarding and rearranging in accordance with the operational rules.

At the present time I have a glossary of one hundred and sixty Russian words with corresponding English equivalents, all systematically coded. The result of this is a coded and typed glossary of approximately one thousand items and fifteen operational rules.

At the very beginning I decided to work on "unsplit" lexical items. *I* have done this not exactly by preference, but because it is quite obvious that "split" or "unsplit" words of the source language would still present the same problem, namely:

a. translation - the proper correspondence of source lexical items to lexical items of the target language.

b. ambiguity.

c. selection.

Since my aim was to develop a code system and operational rules by which I could achieve an automatic process of selection by means of matching code symbols and at the same time, with the same process, to solve the ambiguities of the source language, it was obvious to me that it would be easier to do that by analysing the lexical items as whole items and in context and to postpone the decision as to what would be more economical for the purpose of the machine (split or unsplit) to a later date when the problems of actual translation would be answered.

Thus, as I said above, I have listed in the source and target language glossaries lexical items with all affixes of declension or conjugation etc. This does not mean, however, that the glossaries will have to remain this way. The question of "split" or "unsplit" is absolutely irrelevant to the method I am proposing - by introducing additional operational rules I can get the constituents of my code from the split items. The only advantage of not splitting is that it permits one to list the words as complete items, whereas with splitting the code constituents would be split in the same manner as the lexical items. Otherwise the process will remain the same.

EXAMPLES OF CARDS (GLOSSARY ENTRIES) IN SOURCE AND TARGET LANGUAGES.

ВЗАИМОДЕЙСТВИЯ L2g2221 Slcgl011
Slcpil0ll
Slcpv1011

L2g2221 Slcgl0ll
of interaction

Slcgl0ll = slcil0ll

Slcil0ll
interacticn

Slcpil0ll
interactions

Slcpvl0ll = Slcpil0ll

REZUL'TATY	IZUCHENIIA	OSAEKOV	OSNOVNYKH
Results(N.nom.pl.)	of study (N.gen.sing.)	of precipitates	of basic (A.gen.pl.)
Results (N.acc.pl.)	studies (N.nom.pl.)	(N.gen.pl.)	
	studies (N.acc.pl.)		

SOLEI	POLUCHENNYKH	OT	VZAIMODEISTVIIA
of salts (N.gen.pl.)	of received (PA.gen.pl.)	from (P.gen.)	of interaction
X	received (PA.loc.pl.)		(N.gen.sing.)
			interactions (N. nom.pl.)
			interactions (N.acc pl.)

SUL'FATA	OKISNOGO	ZHEIEZA	S	RAZLICHNYMI
of sulfate X	of oxide X	of ferrum X	with (P.instr.)	with various
(N.gen.sing.)	(A.gen.sing.)	(N.gen.sing.)	for (P.acc.)	(A.instr.pl.)
			from (P.gen.)	

SHCHELOCHAMI,	PRIVELI	K	ZNACHITEL'NOMU	CHISLU
with alkalies X	led (V.pl. demands	to (P.dat.)	to considerable	to number
(N.instr.pl.)	dat,acc, instr.)		(A.dat.sing.)	(N.dat.sing.)

SOEDINENII	S	MENIAIUSHCHIMSIA	OTNOSHENIEM	MEZHDU
of compounds	with (P. instr.)	with varying	with ratio X	between (P.instr.
(N.gen.pl.)	for (P.acc.)	(PA.instr. sing.)	with relation	among (P.gen.)
	from (P.gen.)	(N.instr.sing.)		

OKIS'IU	ZHELEZA	I	SERNYM	ANGIDRIDOM
with oxide X	with ferrum X	and	with	with anhydride X (S.instr. sing.)
(N.instr.sing.)	(N.instr.sing.)	(C,)	sulfur X	
			(A.instr.sing.)	with trioxide X (S.instr. sing.)

OKISNOGO ZHELEZA	=	of ferric oxide
OKIS'IU ZHELEZA	=	with ferric oxide
SERNYM ANGIDRIDOM	=	with sulfur trioxide

Results of study of precipitates of basic salts, received from interaction of sulfate of ferric oxide with various alkalies, led to considerable number of compounds with varying ratio between ferric oxide and sulfur trioxide.

SAMPLE

```
КИНЕТИКА        ГИДРОЛИЗА        СОЛЕЙ                ЖЕПЕЗА
Slcil222        L2g2221+Slcgl221 L2g2221+SlcpglllIx   L2g2221+Slcgl200x
Kinetics        of hydrolysis    of salts             of ferrum

                                         L2gr2221+A.3pg3311-Slcpgllllx        •
                                          of ferric salts


И        УСЛОВИЯ            ОБРАЗОВАНИЯ             ОСНОВНЫХ
C9992   L2g2221+Slcgl312    L2g2221+Slcgl00ly       L2g2221^3pc3322
and        of condition      of formation               of basic
         Slcpil312          Slcpil001y              A3pr3322
         conditions         formations              basic
         Slcpvl312          SlcpvlOOly
         conditions         formations
                            L2g2221+Slcgl444
                               of education


СОЛЕЙ                 ИЗУЧЕНЫ                        НЕДОСТАТОЧНО,
 L2g2221+Slcpgll11x  T5553+T5535+V5252p             N4244
    of salts          have been studied             insufficiently
```