dr inż. Mirosław Gajer
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Katedra Automatyki

THE IMPLEMENTATION OF THE PATTERN-BASED MACHINE TRANSLATION TECHNIQUE FOR THE POLISH LANGUAGE

The high-quality machine translation between natural languages has for a long time been an unattainable dream for many computer scientists involved in this fascinating and interdisciplinary field of the application of computers. The developed quite recently case-based machine translation technique seems to be a serious alternative to the other existing automatic translation methods. In the paper the implementation of the case-based machine translation technique for the development of a system, which would be able to translate an unrestricted foreign texts into Polish is proposed. The new approach to the case-based machine translation technique that takes into account the peculiarity of the Polish grammar is proposed. This new approach was called by this author the pattern-based machine translation techniques because it is based on translation patterns that imply the rules of the grammar of the Polish language. The obtained primary results of the translation experiments conducted for several different languages seem to be very promising and this author believes that they are a step made in the right direction towards a construction of a fully-automatic high-quality machine translation system for translation of unrestricted texts into the Polish language.

1. INTRODUCTION

Machine translation is a science about writing computer programs that are able to translate in a fully automatic way between natural languages, e.g. between Spanish and Korean [1]. Machine translation was invented by an American Warren Weaver, who on the 15th of July 1949 sent his famous memorandum to Rockefeller Foundation with which he started the scientific research in the field [2]. The first research group was constituted on the MIT in 1952. The first public demo of an operating machine translation system took place in New York on the 7th January 1954 [8]. During this demo the machine translation system translated 49 simple sentences from Russian into English.

But, after over 50 year of intensive scientific research machine translation is still far away from its satisfactory solution [4].

This paper describes the implementation of the case-based machine translation technique for the Polish language. At present the case-based machine translation seems to be a very serious alternative to the other translation methods existing so far, and the results obtained by the usage of this technique are generally better (especially for general language) than these of other, e.g. interlingua or transfer approaches.

The paper is organized as follows. In Section 2 it is explained why the automation of the process of translation between natural languages is such a hard task. In Section 3 this

author's proposition of implementation the case-based machine translation technique for the Polish language is described in detail. In Section 4 the clue ideas of pattern-based translation approach are presented. The paper is concluded in Section 5.

2. WHY AUTOMATION OF TRANSLATION IS A HARD TASK?

There are many reasons why machine translation is so extremely difficult. One of them is the ambiguity of natural language. The ambiguity can manifest on the word level (words can have many different equivalents in the target language), syntactic level (the syntactic analysis of a given sentence can be performed in many different ways) and semantic level (the whole sentence can be understood in many different ways).

A very good example of lexical ambiguity is an English word "case" which can have many potential Polish equivalents depending on its context:

1. wypadek, 2. przypadek chorobowy, 3. sprawa sądowa, 4. argumenty przemawiające za czymś, 5. przypadek gramatyczny, 6. stan, 7. sytuacja, 8. warunki, 9. skrzynia, 10. paka, 11. pudełko, 12. szkatułka, 13. futerał, 14. etui, 15. kasetka, 16. koperta zegarka, 17. okno inspektowe, 18. gablotka, 19. torebka nasienna, 20. neseser, 21. pokrowiec, 22. papierośnica

Which one of these 22 possible equivalents should the machine translation system choose in order to produce the correct translation? And how can one automate the inference processes that are necessary for achieving this goal? These, although not beyond any reach and cognition, are still open questions of machine translation science.

A good example of syntactic ambiguity is the following sentence:

I see a man in a park with a telescope.

Should the phrase "with a telescope" be connected with a noun "a man", with a noun "a park" or with the verb "to see"? Depending on these alternative possibilities one can obtain three different Polish translations of this English sentence:

Widzę w parku człowieka z teleskopem.
Widzę człowieka w parku, w którym jest teleskop.
Widzę człowieka w parku za pomocą teleskopu.

Ambiguity of the natural language can also manifest on the highest semantic level of a sentence. One of possible examples is the following sentence:

Cleaning fluids can be very expensive.

Reading such sentence and knowing nothing about the fluids mentioned and the process of cleaning one can not even understand this sentence (i.e., what is really expensive the cleaning process or the fluids itself?). Thus even a human translator would not know which of the Polish translation is the correct one and which one to choose:

Czyszczenie płynów może być bardzo drogie.
Płyny czyszczące mogą być bardzo drogie.

The other reason of fundamental difficulties with translation process are the differences between the grammar systems of source and target languages. Also very important are differences in the vocabularies of source and target languages, where it is often difficult to find the exact equivalents of source words. Last but not least, is that many elements of the source text must be translated in an idiomatic way.

A several year ago case-based or example-based machine translation technique emerged as a very serious alternative to other methods used so far for the automating of translation process [3], [5], [6], [7].

The main idea of CBMT (Case-Base Machine Translation) is as follows. Let it be given any bilingual text. For example the following English text:

There are certain themes of which the interest is all-absorbing, but which are too entirely horrible for the purposes of legitimate fiction. These the mere romanticist must eschew, if he do not wish to offend, or to disgust. They are with propriety handled only when the severity and majesty of truth sanctify and sustain them.

and Polish text:

Istnieją pewne tematy przejmujące do głębi, wszelako nazbyt okropne, izby mogły stać się przedmiotem rzetelnej literatury powieściowej. Rzetelni pisarze muszą ich unikać, jeśli nie chcą wywołać zgorszenia lub niesmaku. Godzi się je podejmować tylko w takich razach, gdy uświęca je i usprawiedliwia surowe dostojeństwo prawdy.

The upper text is the beginning of the story by Edgar Allan Poe entitled "The Premature Burial" and the lower one is its Polish translation made by a prominent Polish translator Stanisław Wyrzykowski.
In this bilingual text the English phrase "they are with proprietary handled" was translated into Polish by a human professional translator as "godzi się je podejmować". So the idea of the CBMT is such that, while translating any other English text into Polish, if the English phrase "they are with proprietary handled" appears anywhere in the translated text once again and it will be substituted the same way as it had been done before by a human translator, i.e. it will be substituted by "godzi się je podejmować" , the translation obtained in such a manner should be still correct.
Of course, there is no certainty that the translation obtained by the use of the CBMT will be always and under any circumstances correct. The main reason for this is a natural language ambiguity, which is a permanent feature of any human language. Even translations of whole sentences can be totally different, depending on the context in which these sentences are used. A good example for this is a Dutch sentence:

Als ik in Amsterdam werkte kopte ik en auto.

This sentence can be translated into English as:

If I worked in Amsterdam I would buy a car.

or as:

When I worked in Amsterdam I bought a car.

Everything depends on the broader context, in which the sentence is used. So, one can see that the CBMT is not always able to return a correct translation, but the probability that the translation will be correct is still very high, and generally the quality of a text translated with the use of the CBMT is much higher than one obtained with the use of other, e.g. transfer or interlingua, methods [9].

3. THE IMPLEMENTATION OF CBMT FOR POLISH

The Polish language, which belongs to the group of Slavonic languages posses a quite complex grammar. The main feature of Polish is its flexion nature, which makes learning it extremely difficult for foreigner. In the Polish grammar there can be found such phenomena as declination of nouns (there are 6 cases) and conjugation of verbs. Also

there are singular and plural forms of words and 3 genders (masculine, feminine and neutral).

These phenomena make the direct usage of the CBMT for Polish language difficult and non-effective. In order to implement the CBMT for the Polish language this translation technique must be modified thoroughly. This author proposed such modification, which takes into account the peculiarities of the Polish grammar and that allows for it affective implementation for the purpose of translating into Polish from the most of West European languages.

In the system developed by this author there exist 3 different types of translation examples, which are further called translation pattern. The whole translation method of using translation patterns was called by this author the Pattern-Based Machine Translation (PBMT). The way in which the translation patterns are translated into Polish depends on the actual value of 3 attributes: <case>, <number>, and <gender>.

The first type of translation pattern is the so called the declination-type translation pattern, which has the following form:

"source language phrase"	< case >
"translation into target language for <case>=1"	1
"translation into target language for <case>=2"	2
"translation into target language for <case>=3"	3
"translation into target language for <case>=4"	4
"translation into target language for <case>=5"	5
"translation into target language for <case>=6"	6
< case > = x; < number > = y; < gender > = z;	

The second type of translation pattern is the conjugation-type translation pattern:

"source language text"	< number >	< gender >
"translation into target language for <number>=1 and <gender>=1"	1	1
"translation into target language for <number>=1 and <gender>=2"	1	2
"translation into target language for <number>=1 and <gender>=3"	1	3
"translation into target language for <number>=2 and <gender>=1"	2	1
"translation into target language for <number>=2 and <gender>=2"	2	2
"translation into target language for <number>=2 and <gender>=3"	2	3
< case > = x; < number > = y; < gender > = z;		

The third type of translation pattern is the non-flexion-type translation pattern:

"source language text"
"translation into target language"
< case > = x; < number > = y; < gender > = z;

As one can see each of the translation patterns has in its bottom the optional information about the values that must be assigned to <case>, <number>, and <gender> attributes, so as the grammatical relations in Polish text between subject, verb, and object should be preserved. However, this part of the translation pattern is optional, i.e. not all the attributes must be assigned values in every translation pattern. In some special cases no attributes can be assigned values at all. The <case> attribute can be assigned values from 1 to 6 (depending on the grammatical case). The attribute <number> is assigned value 1 for singular or 2 for plural. And the attribute <gender> is assigned value 1 for masculine, 2 for feminine, or 3 for neutral.
The manner in which the machine translation system proposed by this author operates will be describe thoroughly in the next paragraph.

4. PATTERN-BASED MACHINE TRANSLATION TECHNIQUE

The way in which the system proposed by this author operates is illustrated on the following example. Let it be given the following sentence [8]. The aim of the pattern-based machine translation system is to translate this sentence into Polish:

Weaver's memorandum brought to the attention of a wide circle the possibilities of a new and exciting application of the computers whose potentialities were being discovered and proclaimed with enthusiasm and optimism at this time.

First at the beginning of a new sentence, which is to be translated the value of the attribute <case> is always set to 1, because the subject in Polish is always in a nominative case.
The translation patterns are taken from the database in the following manner:

1) the declination-type translation pattern

Weaver's memorandum	< case >
wezwanie Weavera	1
wezwania Weavera	2
wezwaniu Weavera	3
wezwanie Weavera	4
wezwaniem Weavera	5
wezwaniu Weavera	6
< number > = 1; < gender > = 3;	

2) the conjugation-type translation pattern

brought to the attention of a wide circle	< number >	< gender >
zwrócił uwagę szerszych kręgów na	1	1
zwróciła uwagę szerszych kręgów na	1	2
zwróciło uwagę szerszych kręgów na	1	3
zwrócili uwagę szerszych kręgów na	2	1
zwróciły uwagę szerszych kręgów na	2	2
zwróciły uwagę szerszych kręgów na	2	3
< case > = 4;		

3) the declination-type translation pattern

the possibilities	< case >
możliwości	1
możliwości	2
możliwościom	3
możliwości	4
możliwościami	5
możliwościach	6
< number > = 2; < gender > = 3;	

4) the non-flexion-type translation pattern

of
< case > = 2;

(In this translation pattern the only English word "of" is not translated at all, because it has not its direct equivalent in Polish. Here, only the value of the attribute <case> is set to 2.)

5) the declination-type translation pattern

a new and exciting application	< case >
nowe i ekscytujące zastosowanie	1
nowego i ekscytującego zastosowania	2
nowemu i ekscytującemu zastosowaniu	3
nowe i ekscytujące zastosowanie	4
nowym i ekscytującym zastosowaniem	5
nowym i ekscytującym zastosowaniu	6
< number > = 1; < gender > = 3;	

6) the non-flexion-type translation pattern

of
< case > = 2;

7) the declination-type translation pattern

the computers	< case >
komputery	1
komputerów	2
komputerom	3
komputery	4
komputerami	5
komputerach	6
< number > = 2;	
< gender > = 3;	

8) the conjugation-type translation pattern

whose potentialities	< number >	< gender >
którego możliwości	1	1
której możliwości	1	2
którego możliwości	1	3
których możliwości	2	1
których możliwości	2	2
których możliwości	2	3
< number > = 2;		
< gender > = 3;		

9) the conjugation-type translation pattern

were being discovered and proclaimed	< number >	< gender >
został odkryty i ogłoszony	1	1
została odkryta i ogłoszona	1	2
zostało odkryte i ogłoszone	1	3
zostali odkryci i ogłoszeni	2	1
zostały odkryte i ogłoszone	2	2
zostały odkryte i ogłoszone	2	3

10) the non-flexion-type translation pattern

with
z
< case > = 5;

11) the declination-type translation pattern

enthusiasm	< case >
entuzjazm	1
entuzjazmu	2
entuzjazmowi	3
entuzjazm	4
entuzjazmem	5
entuzjaźmie	6
< number > = 1; < gender > = 1;	

12) the non-flexion-type translation pattern

and
i

13) the declination-type translation pattern

optimism	< case >
optymizm	1
optymizmu	2
optymizmowi	3
optymizm	4
optymizmem	5
optymiźmie	6
< number > = 1; < gender > = 1;	

14) the non-flexion-type translation pattern

at this time
w owym czasie

After putting all above translation patterns together the Polish translation of the English sentence takes the following form:

Wezwanie Weavera zwróciło uwagę szerszych kręgów na możliwości nowego i ekscytującego zastosowania komputerów których możliwości zostały odkryte i ogłoszone z entuzjazmem i optymizmem w owym czasie.

It must be noticed that the obtained Polish translation is totally correct from the syntactic point of view. It is also the exact translation of the original English sentence. And last but not least, the Polish translation sounds very natural for its reader, resembling the effect of the work of a human translator. In fact, it would be rather hard to say with a certainty that this translation is only a product made by a dehumanized machine.

5. CONCLUSIONS

Machine translation has now more than 50 years long history of scientific investigation, but the full-automatic high-quality machine translation system is still the Holy Grail of scientific research that maybe will never be found [4]. So, maybe the computers will never eliminate the necessity of employment of the human translator, but the scientist still try to approach as far as possible to the ideal solution in which the high-quality translation of text of a general language can be obtained automatically by the use of a personal computer.

In the earlier epoch of machine translation research it was taken for certain that the deep syntactic and semantic analysis of the text being translated is absolutely necessary if one wants to obtain a perfect translation of the source text. The proposed several years ago example (or case) based machine translation technique revealed that high-quality automatic translation is possible without the complex semantic and syntactic analysis of the translated text [9].

The case-based machine translation technique has never been implemented for the Polish languages. This author's investigation has revealed that the direct implementation of this technique for a Slavonic language such as Polish would be difficult and probably wouldn't bring the desired effects. So, this author modified the case-based machine translation technique a little, which allowed for its implementation for the Polish languages. The modified case-based machine translation technique was called by this author the pattern-based machine translation, because the clue of this translation method are translation patterns that allow for taking into account the flexion nature of the Polish language.

This author has experimented with several languages. First experiments were conducted for other Slavonic languages, which have similar grammatical structure to Polish. The experimental machine translation systems were created for such Slavonic languages as: Russian, Ukrainian, Czech, Slovakian and Croatian. Further experiments were conducted for such West-European languages as: English, German, Dutch, Swedish, Norwegian, Danish, French, Italian, Spanish and Portuguese. This author has also experimented with Greek and Arabic languages. In all case a quite big databases of translation patterns were developed and the obtained primary results look very promising, showing that the proposed by this author pattern-based machine translation technique can lead to the construction of a high-quality machine translation system that would be able to translate from many different languages into the Polish language.

6. REFERENCES

[1] D. Arnold, L. Balkan, S. Meijer, R. L. Humphreys, L. Sadler: *Machine Translation: An Introductory Guide*, NCC Blackwell, London, 1994

[2] P. Whitelock, K. Kilby: Linguistic and Computational Techniques in Machine Translation System Design, UCL Press, London, GB, 1995

[3] R. Perez: *From Novelty to Ubiquity: Computers and Translation at the Close of Industrial Age*, http:\\www.accurapid.com/journal/15mt2.htm

[4] A. Melby: *Machine Translation and Philosophy of Language*; Machine Translation Review, No. 9, April 1999, pp. 6-17

[5] R. D. Brown: *Adding linguistic knowledge to a lexical example-based translation system*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, 2000

[6] R. D. Brown: *Example-based machine translation in the Pangloss system*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, 2000

[7] Y. Zhang, R. D. Brown, R. E. Frederking: *Adapting an example-based translation system to Chinese*, Carnegie Mellon University, Pittsburgh, PA, USA, 2000

[8] W. J. Hutchins: *Machine translation: past, present, future*, Ellis Horwood Limited, New York, 1986

[9] G. Carbonell, T. Mitamura, E. H. Nyberg: *The KANT perspective: A critique of pure transfer(and pure interlingua, pure statistics,...)*, Center for Machine Translation, Carnegie Mellon, University, Pittsburgh, PA, USA, 1998