INFORMATICA, 2002, Vol. 13, No. 4, 417–440

The Implementation of the Example-Based Machine Translation Technique for German-to-Polish Automatic Translation System

Mirosław GAJER
Technical University in Cracow, Department of Control Science
Al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: mgajer@ia.agh.edu.pl

Received: September 2002

Abstract. High-quality machine translation between human 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 example-based machine translation technique seems to be a serious alternative to the existing automatic translation techniques. In the paper the usage of the example based machine translation technique for the development of system, which would be able to translate an unrestricted German text into Polish is proposed. The new approach to the example-based machine translation technique that takes into account the peculiarity of the Polish grammar is developed. The obtained primary results of the development of proposed system seem to be very promising and appear to be a step made in the right direction towards a fully-automatic high quality German-into-Polish machine translation system for unrestricted text.

Key words: natural language processing, computational linguistics, machine translation.

1. Introduction

Machine translation is a science that delivers the knowledge how to program the computers, so as they were able to translate between human languages, for example, between Danish and Bulgarian. It may be amazing, but the field of machine translation is almost as old as the invention of computer itself (Blekhman and Pevzner, 2000). In 1949 an American scientist Warren Weaver sent the memorandum to The Rockefeller Foundation (American institution supporting the scientific research), in which he demanded starting the research on the automation of translation between natural languages (Arnold *et al.*, 1994). Warren Weaver was inspired by cryptographic techniques, which were developed very strongly during the years of The Second World War, and he thought that there existed some fundamental similarities between these cryptographic techniques and the process of translation between human languages (Waibel *et al.*, 2000).

This author doesn't know if Warren Weaver had a good command of any foreign language, but it seems to be clear that the level of Warren Weaver's general linguistic knowledge was rather low. Indeed, it soon appeared that the problem of machine translation is far more complicated and far more harder than Warren Weaver had ever imagined.

2. A Bit of History of the Development of Machine Translation Systems

The first research group dedicated especially to machine translation was established in the United States in 1951. The first public demo of an operating machine translation system was given also in the USA in 1954. During this demo the system translated 49 pre-selected sentences from Russian into English. The system was using a vocabulary of 250 words and only six simple grammatical rules. The possibilities of early machine translation systems were very far from this, what had been expected, and many scientists connected with the field started to be disappointed. In 1966 the ALPAC (Automatic Language processing Advisory Committee) published its famous report, concluding that machine translation was slower, less precise, and more expensive than human translation. As a result, funding of this type of projects was cut (Rico, 1998). The renaissance came in the late 1970s. The United States Air Force funded work on the METAL system at the University of Texas in Austin, and the results of the work of TAUM group led finally to the installation of the METEO system, which was a great commercial success. It is worth to notice that the METEO machine translation system is still in use, and it translates every day from English into French more than 50,000 words of weather forecast bulletins.

Now, a still growing interest of machine translation systems can be observed in many countries, especially in Japan, the United States, the European Union, and India (Bandyopadhyay, 2000), but after so many years of an intensive scientific research high-quality machine translation between human languages for unrestricted text is still a long-term scientific dream of enormous political, social, and scientific importance (Mitamura, 1998). Machine translation was also one of the earliest applications suggested for the computers, but turning this scientific dream into reality has turned out to be much harder, and much more interesting than it had first appeared (Arnold *et al.*, 1994). So, in the next point we will try to give the answer why the automation of translation between natural languages is so difficult?

3. Why Machine Translation is such a Hard Task?

To answer the question about the origin of difficulties with automation of translation between human languages, let us consider the differences which we can discover when we compare some of the human languages.

First of all, when we study grammatical systems of any natural languages that are not closely related with each other, we easily can see that there exist much more differences than similarities between them (Zue and Glass, 2000).

For example let us compare the systems of personal pronouns of Arabic and Hungarian languages.

Personal pronouns system of Hungarian

Singular	Plural
1. *én*	1. *mi*
2. *te*	2. *ti*
3. *ö*	3. *ök*

Personal pronouns system of Arabic:

Singular	Double	Plural
1. *ana*	1. *nahnu*	1. *nahnu*
2. (m.) *anta*	2. *antuma*	2. (m.) *antum*
2. (f.) *anti*		2. (f.) *antunna*
3. (m.) *huua*	3. *huma*	3. (m.) *hum*
3. (f.) *hija*		3. (f.) *hunna*

It's clear that personal pronouns system of Arabic is much more complicated than this one of Hungarian. It is caused by the fact that Hungarian language doesn't know such invention as grammatical gender of the words. Also grammatical number in Hungarian can be only singular or plural, whereas in Arabic it can be singular, plural, or double.

So, one can easily see that translating Hungarian personal pronouns into their Arabic equivalents is a hard task. For example, if we want to translate Hungarian pronoun *ök* (in English *they*) into Arabic we must additionally know how many persons are involved with this pronoun *ök*. If exactly two persons are considered we will use the Arabic word *huma*. But, if there are more than two persons we must additionally know, whether they are men or women. If they are men we will use the Arabic word *hum*, in other case *hunna*.

Where do we know from how many persons are involved, and whether they are men or women, while Hungarian word *ök* states nothing about it? The answer is that we know this from the context of the utterance. A human translator can in most cases very easily extract such context information, but the full automation of this process is still a pure science-fiction.

A quite big differences between human languages can also be noticed when we study their vocabularies. In fact, the vocabulary of each language is an independent and very compound system. If we want to translate, for example, from Chinese into Croatian it's a hard work to find in Croatian the equivalents of Chinese words that preserve their original meanings. While doing that a human translator have to cope with a enormous number of lexical holes, it is, words that don't have their equivalents in other language, and as such can be translated only by the medium of a long description that clears up their semantics.

This situation is illustrated in Fig. 1. In Fig. 1 each rectangle is a symbol of some physically existing object or some abstract entity. The rectangles are numerated from 1 to 6. Further, we have two different natural languages: language A and language B.

We can see that in language A objects 1 and 2 are described only by one common lexical entity, whereas in language B there exist two different lexical entities, separate for object 1 and object 2.

Further, we can notice that the object 3 doesn't have any lexical entity in language A, so it is a lexical hole, whereas in the language B it has its own lexical item.

Objects 4 and 5 in language A are grouped together in one lexical entity and object 6 is a separate lexical entity, while in language B it is otherwise. We can notice that objects 5 and 6 form one lexical entity.

A very good example of this (maybe a bit too abstract) divagation comes from Swedish language. If we want to translate English word *grandfather* into Swedish, we

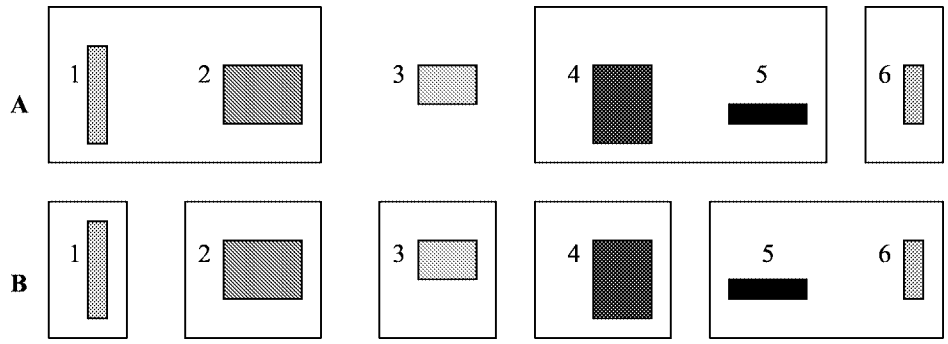

Fig. 1. The illustration of the way in which different languages divide the reality into lexical items.

must additionally know whether this grandfather is a father of a father or a father of a mother. In the first case we should use the Swedish word *farfar* in the other *morfar*, which is illustrated in Fig. 2.

Another similar example comes from French language. Namely, if we want to translate an English word *river* into French, we must know whether it is a main river, which is directly connected with the see, or it is only a tributary of some bigger river. We absolutely must have this information because in the first case we have to use French word *fleuve*, and in the second the correct choice is French word *rivièr*. This situation is illustrated in Fig. 3.

But the perhaps most serious problem, which the computer has to cope with in machine translation is the ambiguity of any human language (Baker *et al.*, 1998). We can distinguish syntactic ambiguity when there exist at least two alternative ways of syntactic analysis of a sentence. One of the examples is an English sentence:

I see a man in the park with a telescope.

This sentence is threefold ambiguous because we don't know if the phrase *with the telescope* should be interpreted in connection with a verb *to see*, or with the noun *a man*, or with the noun *the park*. In each of these cases the meaning of the sentence is totally

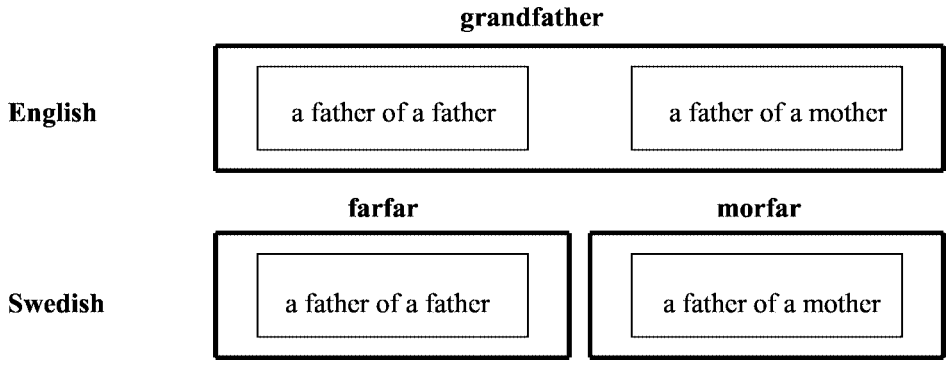

Fig. 2. An English word *grandfather* versus Swedish words *farfar* and *morfar*.

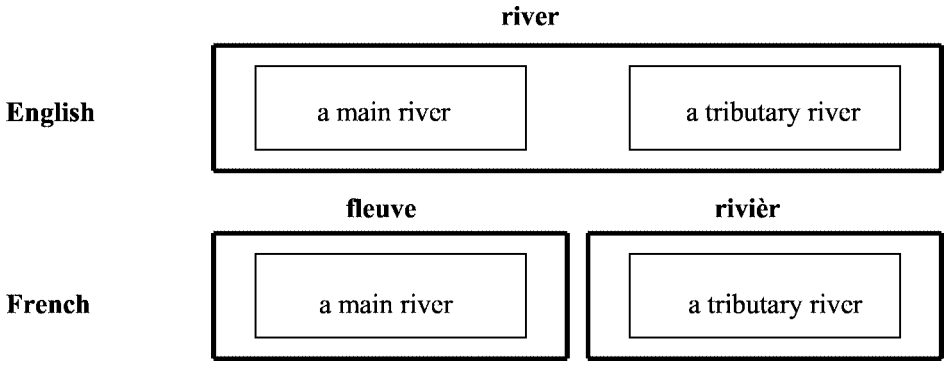

Fig. 3. An English word *river* versus French words *fleuve* and *rivièr*.

different. The problem is that this ambiguity cannot be preserved during the translation because in order to translate, one have to understand the sentence being translated. For example, in the case of translating this sentence into Polish three different translations are possible, depending on the interpretation of the English sentence:

Widzę człowieka w parku za pomocą teleskopu.
Widzę w parku człowieka z teleskopem.
Widzę człowieka w parku z teleskopem.

Another example of the ambiguity on the syntactic level is an English phrase:

old man and women

Analyzing this phrase we don't know whether it is:

old man and old woman

or

old man and woman at any age

Another kind of ambiguity is at the semantic level. Semantic ambiguity appears when one sentence can be understood in at least two different manners (Whitelock and Kilby, 1995). A good example is an English sentence:

She threw the vase at the window and it broke.

This sentence is ambiguous because we don't know what is broken, *a window* or *a vase*? If we would like to try to translate this sentence into Polish we would have a hard choice to make. If we decided that the window is broken, the Polish translation would be:

Ona rzuciła wazą w okno i ono pękło.

In the other case we would obtain the following Polish translation:

Ona rzuciła wazą w okno i ona pękła.

Another kind of ambiguity is the ambiguity at the lexical level of language analysis. Lexical ambiguity is such a serious problem in the case of machine translation systems because it exists in every natural language and it is really ubiquitous. In fact, if we open any bilingual dictionary, for example *The Great English-Polish Dictionary*, it's very hard to find a word that would have only exactly one meaning. In fact, most of English words have at least two completely different Polish equivalents. So, the question is, which one of them the computer should choose while translating, and where can computer know from, which one of them is the correct one?

Let us suppose that we have a sentence built from ten different word, and let each of these words has exactly two different meanings. If the computer chose the equivalents of these words at random this sentence could be translated in 1024 different ways. The probability that acting this way we obtain a correct translation of a whole document built from many such sentences is equal to zero in practice. Moreover, no efficient algorithm that allows for solving this problem is known, and lexical ambiguity can be found in abundance in any human language – bellow are listed some examples of possible Polish translations of lexically ambiguous words taken from several languages of the world.

Polish equivalents of French word **perle** are: 1. perła; 2. paciorek; 3. kapsułka.

Polish equivalents of Spanish word **fondo** are: 1. dno; 2. głębia; 3. tło.

Polish equivalents of Italian word **stufa** are: 1. piec; 2. cieplarnia.

Polish equivalents of German word **Absatz** are: 1. ustęp; 2. obcas; 3. osad; 4. złoże; 5. osadzenie; 6. zbyt.

Polish equivalents of Englis word **butt** are: 1. beczka; 2. pień; 3. pniak; 4. grubszy koniec; 5. kolba karabinu; 6. płastuga; 7. nasyp za strzelnicą; 8. pośmiewisko; 9. uderzenie głową.

Polish equivalents of Dutch word **boodschap** are: 1. poselstwo; 2. polecenie; 3. wiadomość; 4. zakupy.

Polish equivalents of Swedis word **tomten** are: 1. parcela; 2. plac; 3. krasnoludek.

Polish equivalents of Norwegian word **hytte** are: 1. chata; 2. szałas; 3. buda; 4. huta; 5. kabina.

Polish equivalents of Danish word **løber** are: 1. biegacz; 2. dywanik.

Polish equivalents of Finnish word **kanta** are: 1. podstawa; 2. obcas; 3. stanowisko; 4. baza.

Polish equivalents of Greek word σκοπς (*skopos*) are: 1. zamiar; 2. melodia; 3. wartownik.

Polish equivalents of Arabic word وصل (*wusal*) are: 1. połączenie; 2. łącze; 3. kontakt; 4. związek; 5. zawias; 6. dodatek.

Very problematic for machine translation systems are also complex nominal groups like for example:

adult toy manufacturer

This nominal group can be understood as:

manufacturer of toys for adults

or

an adult manufacturer of toys

Another kind of difficulties making the automation of translation process so hard are all idiomatic phrases. The problem is that these idiomatic phrases can be also interpreted literary. For example, an idiom taken for the Hausa language:

Gari ya yi kyau.

means that:

It is a beautiful weather.

But when we treat this sentence literary it means that:

The town is beautiful.

So, which of this two meanings is the correct one? A human translator basing on the context of this idiom can probably make the right decision, but the automation of such inference is still far beyond the possibilities of any computer system.

4. Is High-Quality Machine Translation Possible?

Taking under consideration all above mentioned factors translation between natural languages can be seen as a highly creative process. A human translator must have a lot of invention and must know how to deal with the situations he had never met before. So, the right question is, whether it is possible to replace a human by a computer?

A prominent physicist Roger Penrose in his famous books on artificial intelligence, entitled "New Caesar's Mind" (Penrose, 1995) and "The Shadows of the Mind" gave very strong arguments supporting his thesis that the human brain operates in a non-algorithmic manner and because of this fact a human mind cannot be fully simulated by computer.

Thus, if we cannot replace a human by a computer does it also mean that a fully-automatic high-quality machine translation for unrestricted text is impossible (Fukutomi, 2000; Murphy, 2000; Nyberg *et al.*, 1998; Mitamura, 1998)?

A philosopher Alan Melby in his paper (Melby, 1999) states that machine translation is headed in the right direction as far as domain-specific approaches using controlled languages are concerned. But further work on fully-automatic high-quality machine translation of unrestricted text is a waste of time and money. To build such systems a real breakthrough in natural language processing (and maybe in the whole filed of information processing) is required. Moreover, such breakthrough will not be based on any extension of currently known techniques, as electric bulb was not invented just because the research on the candle had been conducted (Melby, 1999).

5. Example-Based Machine Translation Technique

The arguments given by Roger Penrose are very strong and it's not possible to ignore them any further. So, probably Alan Melby is right that replacing a human translator totally is not possible basing only on the currently known techniques. But, by using these currently known techniques we can still try to approach as close as possible to this unattainable goal, which is a fully-automatic high-quality machine translation for unrestricted text. Suppose that during the intensive scientific research we built a machine translation system, which gives a translation of 99% accuracy, while operating on an unrestricted text (only 1% of this text need to be approved by a human translator). So can we really say, like Alan Melby, that we had wasted time and money on this research?

Up till now, many totally different approaches to machine translations have been developed. These are, among others: syntactic transfer, interlingua-based machine translation, knowledge-based machine translation, systems based on statistics or neural nets, etc. (Ney *et al.*, 2000; Canals *et al.*, 2000; Loukachevitch and Dobrov, 2000). Among

these example-based machine translation is becoming a serious alternative paradigm, but in most cases it is still an unproven technique, which is in its early research phase (Carbonell *et al.*, 1998).

But it is not always so. One prominent example comes from Spain. The case of the magazine entitled Periódico de Catalunya is interesting because it is probably the first fully operational machine translation system for translation of unrestricted text that has ever been built, which produces nearly hundred percent satisfactory results while translating from Spanish into Catalan. It is really amazing that this machine translation system is not based on any of the currently known computational linguistics theories. Moreover, it does not analyze the sentence in any way it only replaces source words (or groups of words) by their target equivalents, just like spelling-checker would do. The system has a huge dictionary that effectively replaces all linguistic analysis of the source text. The development of the systems requires a lot of work, in fact a quite big team of trained linguists constantly updates the dictionary with new terms, verbs in their different forms and sequences of words of up to six elements. Up till now, it has been probably the only practical implementation of a purely unsophisticated machine translation system basing only on a pattern-matching scheme (Rico, 1998).

So, can this Spanish-Catalan system be an example showing the way how to mysteriously solve the problem of building a fully-automatic high-quality machine translation system?

The answer to this question is not so obvious as somebody may think. We can not omit the fact that this Spanish-Catalan system benefits to the great measure from the similarities of the two languages involved in the machine translation process. In fact, the differences between Spanish and Catalan languages are rather minor and have in most cases only phonological nature and more rarely morphological or grammatical.

The results obtained during the development of Spanish-Catalan machine translation system can be obviously applied to any system, which translates between closely related languages. Probably acting this way an effective machine translation systems for unrestricted text can be built for such pair of languages like Swedish and Norwegian, Norwegian and Danish, Swedish and Danish, Spanish and Portuguese, German and Dutch, or Finnish and Estonian. But it is a bit doubtful if doing this way a high-quality machine translation system for unrestricted text can be built, which would be able to translate between a pair of totally typologically different and genetically unrelated languages like, for example, Chinese and French.

In order to allow the reader to imagine how difficult the translation between unrelated languages is, the results of a following experiment are presented beneath (Majewicz, 1989). It was taken some sample of text written in Polish, the elements (words or phrases) of which were numbered in the following way.

The original Polish text:

Analiza tych dwóch elementów zwyczaju międzynarodowego posiada wielkie znaczenie z uwagi na wyrok Międzynarodowego Trybunału Sprawiedliwości z 29 listopada 1950r. (w

sporze między Boliwią a Peru o prawo azylu), który stwierdza, że państwo, które powołuje się na zwyczaj międzynarodowy według art. 38 b musi przeprowadzić dowód, iż powstałon w sposób wiążący drugie państwo.

The order of words in the Polish text:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 47

Then, this text was translated into English and the order of the English equivalents of the elements of the Polish source text was the following.

The English translation of the text:

The analysis of these two elements of the international custom is of a great importance in view of the sentence of the International Court of Justice of the 29th of November 1950 (concerning the dispute between Bolivia and Peru about the right of asylum) which states that a state that is referring to the international custom quoting Article 38b must present the evidence that the custom emerged in a way confining the other state.

The order of words in the English translation:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, **33**, **32**, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 47

We can see that in the English translation the word order is almost the same as in the Polish original text. Only two elements (32 and 33) are swapped. This is a good notion indicating that maybe example-based machine translation technique can be applied successfully to the pair of English and Polish languages.

Further the same text was translated into Japanese and the word order was the following:

The Japanese translation of the text:

Kokusai kanshu-no kono futatsu-no yoso-no bunseki-wa daisanjuhachi-jo-bi-ni shitagatte kokusai kanshu-o in'yo suru kokka-wa kokusai kan-shu-ga ta-no kokka-o kosoku suru hoho-de sonzai suru to iu shoko-o teishutsu shinakereba naranai to iu koto-o kakunin suru (higoken-ni kansuru boribia peru kan-no funso-ni tsuite-no) senkyuhyakugojunen juichigatsu nijukunichi-no kokusai shiho saibansho-no hanketsu-kara mite hijo-na juyoseimotte iru.

The order of words in the Japanese translation:

6, 5, 2, 3, 4, 1, 36, 35, 37, 34, 33, 32, 30, 28, 43, 46, 47, 45, 44, 42, 40, 39, 38, 26, 24, 23, 22, 19, 21, 21, 18, 17, 16, 15, 12, 14, 13, 11, 10, 8, 9, 7

We can see that the word order in the Japanese translation is totally different than in the Polish source text. It doesn't sound good if we would like to think about using an example-based machine translation technique for such totally unrelated and typologically different languages.

6. Example-Based Machine Translation for the Polish Language

The Polish language, which belongs to the group of the Slavonic languages, differs very much in its grammar from the West-European languages. This is a reason why a direct implementation of the example-based machine translation technique for the Polish language is not so easy and probably wouldn't bring the desired effects. In order to use the example-based machine translation for the system, which translates from West-European languages into Polish, the example-based translation technique must be modified a bit.

The system proposed by this author is based on the following observations:

- Perhaps, in every human language we can distinguish the first, the second, and the third grammatical person.
- Perhaps, in every human language we can distinguish such elements of the sentence as: a subject group S, a verb group V, and an object group O. In the majority of the Indo-European languages the most common word order in the sentence is SVO (subject-verb-object).
- In the Polish language the grammatical person, number, and gender of the verb group must agree with these of the subject group. Further, the grammatical case of the object group must agree with the one required by the verb group.

Taking into account the abovementioned observations the architecture of the propose example-based machine translation system is the following. The system is based on the database in which the translation examples are stored. The database records can have different attributes, such as: <case>, <number>, <person>, and <gender>. During the process of translation the values of these attributes are set respectively to the grammatical rules of the Polish language, so as the subject and verb agreed in the grammatical person, number, and gender. Also the grammatical case of the object must agree with the one required by the verb.

There exist three types of translation examples:

1) Noun group translation examples. These translation examples play the role of the subject or the object of the sentence, which is to be translated. The database record for a noun group translation example has the following form:

NG_source	<case>
NG_target_1	1
NG_target_2	2
NG_target_3	3
NG_target_4	4
NG_target_5	5

NG_target_6	6	
<person>	<number>	<gender>
PERSON	NUMBER	GENDER

- NG_source – a noun group of the source language that is to be translated into the target language (Polish) in a manner depending on the value of the attribute <case>
- NG_target_1 – translation of the source language noun group when <case> = 1
- NG_target_2 – translation of the source language noun group when <case> = 2
- NG_target_3 – translation of the source language noun group when <case> = 3
- NG_target_4 – translation of the source language noun group when <case> = 4
- NG_target_5 – translation of the source language noun group when <case> = 5
- NG_target_6 – translation of the source language noun group when <case> = 6
- PERSON – the value to which the attribute <person> is set, so as the subject and verb groups agreed in their grammatical person
- NUMBER – the value to which the attribute <number> is set, so as the subject and verb groups agreed in their grammatical number
- GENDER – the value to which the attribute <gender> is set, so as the subject and verb groups agreed in their grammatical gender

2) Verb group translation examples. These translation examples play the role of the verb of the sentence, which is to be translated. The data base record for verb group translation examples has the following form:

VG_source	<person>	<number>	<gender>
VG_target_1	1	1	1
VG_target_2	1	1	2
VG_target_3	1	2	1
VG_target_4	1	2	2
VG_target_5	1	1	1
VG_target_6	2	1	2
VG_target_7	2	2	1
VG_target_8	2	2	2
VG_target_9	3	1	1
VG_target_10	3	1	2
VG_target_11	3	1	3
VG_target_12	3	2	1
VG_target_13	3	2	2
VG_target_14	3	2	3
<case>			
CASE			

- VG_source – verb group of the source language that is to be translated into the target language (Polish) in a manner depending on the value of the attributes <person>, <number>, and <gender>

- VG_target_1 – translation of the source language noun group when <person> = 1, <number> = 1, and <gender> = 1
- VG_target_2 – translation of the source language noun group when <person> = 1, <number> = 1, and <gender> = 2
- VG_target_3 – translation of the source language noun group when <person> = 1, <number> = 2, and <gender> = 1
- VG_target_4 – translation of the source language noun group when <person> = 1, <number> = 2, and <gender> = 2
- VG_target_5 – translation of the source language noun group when <person> = 2, <number> = 1, and <gender> = 1
- VG_target_6 – translation of the source language noun group when <person> = 2, <number> = 1, and <gender> = 2
- VG_target_7 – translation of the source language noun group when <person> = 2, <number> = 2, and <gender> = 1
- VG_target_8 – translation of the source language noun group when <person> = 2, <number> = 2, and <gender> = 2
- VG_target_9 – translation of the source language noun group when <person> = 3, <number> = 1, and <gender> = 1
- VG_target_10 – translation of the source language noun group when <person> = 3, <number> = 1, and <gender> = 2
- VG_target_11 – translation of the source language noun group when <person> = 3, <number> = 1, and <gender> = 3
- VG_target_12 – translation of the source language noun group when <person> = 3, <number> = 2, and <gender> = 1
- VG_target_13 – translation of the source language noun group when <person> = 3, <number> = 2, and <gender> =2
- VG_target_14 – translation of the source language noun group when <person> = 3, <number> = 2, and <gender> = 3
- CASE – the value to which the attribute <case> is set, so as the grammatical case of the object agreed with the one required by the verb group

3) Other translation examples. These translation examples are the same as in the classical technique of example based machine translation. In the sentence they can play the role of particles, junctions, etc. The data base record for this kind of translation examples has the following form:

EX_source
EX_target

- EX_source – a phrase of the source language that is to be translated into the target language (Polish)
- EX_target – translation of the source language phrase

7. The German-to-Polish Example-Based Machine Translation System

The proposed machine translation technique was implemented by this author for the system, which translates from German into Polish. The German language belongs to the group of Germanic languages, and it differs much from Polish, which is a Slavonic language. What is important is that both languages belong to the Indo-European family of languages that implies their grammatical structures to be similar enough, so that the example-based machine translation could be used.

This author developed a database of translation examples, according with the proposed by himself methodology, which allowed to translate simple texts from German into Polish. The manner in which the system operates is illustrated on the following example.

The purpose of the proposed system is to translate into Polish the following German text composed of a few simple sentences:

Berlin wurde urkundlich erwähnt zum ersten Mal im Jahre 1244. Berlin war in dieser Zeit ein sehr kleines Dorf. Nach und nach jedoch Berlin wurde immer größer. Im Jahre 1740 Berlin wurde die preußische Residenzstadt. Im Jahre 1871 Berlin erhieltet eine zentrale Bedeutung für Deutschland. Berlin wurde Reichshauptstadt. Berlin hat als Kulturstadt internationalen Ruf. Hier gibt es viele Museen Hochschulen und Theater.

First, at the beginning of the sentence the value of attribute <case> is set to 1, because the subject in the Polish sentence is always in a nominative case.

<case> =1;

The translation examples are taken from the database in the order of their occurrence in the translated sentence:

1) noun group translation example

Berlin		
<case>	1	
Berlin	2	
Berlina	3	
Berlinowi	4	
Berlin	5	
Berlinem	6	
<person>	<number>	<gender>
3	1	1

2) verb group translation example

wurde urkundlich erwähnt	<person>	<number>	<gender>
–	1	1	1
–	1	1	2
–	1	2	1

–	1	2	2
–	2	1	1
–	2	1	2
–	2	2	1
–	2	2	2
zostałwspomniany w dokumentach	3	1	1
została wspomniana w dokumentach	3	1	2
zostało wspomniane w dokumentach	3	1	3
zostali wspomniani w dokumentach	3	2	1
zostały wspomniane w dokumentach	3	2	2
zostały wspomniane w dokumentach	3	2	3

<case>
–

3) translation example

zum ersten Mal
po raz pierwszy

4) translation example

im Jahre 1244
w roku 1244

At the beginning of a new sentence the attribute <case> is set to 1.
<case> = 1;

5) noun group translation example

Berlin	<case>
Berlin	1
Berlina	2
Berlinowi	3
Berlin	4
Berlinem	5
Berlinie	6

<person>	<number>	<gender>
3	1	1

6) verb group translation example

war	<person>	<number>	<gender>
–	1	1	1
–	1	1	2
–	1	2	1
–	1	2	2
–	2	1	1
–	2	1	2
–	2	2	1
–	2	2	2
był	3	1	1
była	3	1	2
było	3	1	3
byli	3	2	1
były	3	2	2
były	3	2	3

<case>
5

7) translation example

in dieser Zeit
w tym czasie

8) noun group translation example

ein sehr kleines Dorf	<case>
bardzo mała wieś	1
bardzo małej wsi	2
bardzo małej wsi	3
bardzo małą wieś	4
bardzo małą wsią	5
bardzo małej wsi	6

<person>	<number>	<gender>
3	1	2

At the beginning of a new sentence the attribute <case> is set to 1.
<case> = 1;

9) translation example

nach und nach
stopniowo

10) translation example

jedoch
jednakże

11) noun group translation example

Berlin	<case>
Berlin	1
Berlina	2
Berlinowi	3
Berlin	4
Berlinem	5
Berlinie	6

<person>	<number>	<gender>
3	1	1

12) verb group translation example

wurde immer größer	<person>	<number>	<gender>
–	1	1	1
–	1	1	2
–	1	2	1
–	1	2	2
–	2	1	1
–	2	1	2
–	2	2	1
–	2	2	2
stawałsię coraz większy	3	1	1
stawała się coraz większa	3	1	2
stawało się coraz większe	3	1	3
stawali się coraz więksi	3	2	1
stawały się coraz większe	3	2	2
stawały się coraz większe	3	2	3

<case>
–

At the beginning of a new sentence the attribute <case> is set to 1.
<case> = 1;

13) noun group translation example

Berlin	<case>
Berlin	1
Berlina	2
Berlinowi	3
Berlin	4

Berlinem	5
Berlinie	6

<person>	<number>	<gender>
3	1	1

14) verb group translation example

wurde	<person>	<number>	<gender>
–	1	1	1
–	1	1	2
–	1	2	1
–	1	2	2
–	2	1	1
–	2	1	2
–	2	2	1
–	2	2	2
został	3	1	1
została	3	1	2
zostało	3	1	3
zostali	3	2	1
zostały	3	2	2
zostały	3	2	3

<case>
5

15) noun group translation example

die preußische Residenzstadt	<case>
stolica Prus	1
stolicy prus	2
stolicy Prus	3
stolicę Prus	4
stolicą Prus	5
stolicy Prus	6

<person>	<number>	<gender>
3	1	2

16) translation example

im Jahre 1740
w roku 1740

At the beginning of a new sentence the attribute <case> is set to 1.
<case> = 1;

17) noun group translation example

Berlin	<case>
Berlin	1
Berlina	2
Berlinowi	3
Berlin	4
Berlinem	5
Berlinie	6

<person>	<number>	<gender>
3	1	1

18) verb group translation example

erhiltet	<person>	<number>	<gender>
–	1	1	1
–	1	1	2
–	1	2	1
–	1	2	2
–	2	1	1
–	2	1	2
–	2	2	1
–	2	2	2
uzyskał	3	1	1
uzyskała	3	1	2
uzyskało	3	1	3
uzyskali	3	2	1
uzyskały	3	2	2
uzyskały	3	2	3

<case>
4

19) noun group translation example

eine zentrale Bedeutung	<case>
centralne znaczenie	1
centralnego znaczenia	2
centralnemu znaczeniu	3
centralne znaczenie	4
centralnym znaczeniem	5
centralnym znaczeniu	6

<person>	<number>	<gender>
3	1	3

20) translation example

für Deutschland
dla Niemiec

21) translation example

im Jahre 1871
w roku 1871

At the beginning of a new sentence the attribute <case> is set to 1.
<case> = 1;

22) noun group translation example

Berlin	<case>
Berlin	1
Berlina	2
Berlinowi	3
Berlin	4
Berlinem	5
Berlinie	6

<person>	<number>	<gender>
3	1	1

23) verb group translation example

wurde	<person>	<number>	<gender>
–	1	1	1
–	1	1	2
–	1	2	1
–	1	2	2
–	2	1	1
–	2	1	2
–	2	2	1
–	2	2	2
został	3	1	1
została	3	1	2
zostało	3	1	3
zostali	3	2	1
zostały	3	2	2
zostały	3	2	3

<case>
5

24) noun group translation example

Reichshauptstadt	<case>
stolica Rzeszy	1
stolicy Rzeszy	2
stolicy Rzeszy	3
stolicę Rzeszy	4
stolicą Rzeszy	5
stolicy Rzeszy	6

<person>	<number>	<gender>
3	1	2

At the beginning of a new sentence the attribute <case> is set to 1.
<case> = 1;

25) noun group translation example

Berlin	<case>
Berlin	1
Berlina	2
Berlinowi	3
Berlin	4
Berlinem	5
Berlinie	6

<person>	<number>	<gender>
3	1	1

26) verb group translation example

hat	<person>	<number>	<gender>
–	1	1	1
–	1	1	2
–	1	2	1
–	1	2	2
–	2	1	1
–	2	1	2
–	2	2	1
–	2	2	2
ma	3	1	1
ma	3	1	2
ma	3	1	3
mają	3	2	1
mają	3	2	2
mają	3	2	3

<case>
4

27) translation example

als Kulturstadt
jako miasto kultury

28) noun group translation example

internationalen Ruf	<case>
sława międzynarodowa	1
sławy międzynarodowej	2
sławie międzynarodowej	3
sławę międzynarodową	4
sławą międzynarodową	5
sławie międzynarodowej	6

<person>	<number>	<gender>
3	1	2

At the beginning of a new sentence the attribute <case> is set to 1.
<case> = 1;

29) translation example

hier
tutaj

30) verb group translation example

gibt es	<person>	<number>	<gender>
znajduje się	1	1	1
znajduje się	1	1	2
znajduje się	1	2	1
znajduje się	1	2	2
znajduje się	2	1	1
znajduje się	2	1	2
znajduje się	2	2	1
znajduje się	2	2	2
znajduje się	3	1	1
znajduje się	3	1	2
znajduje się	3	1	3
znajduje się	3	2	1
znajduje się	3	2	2
znajduje się	3	2	3

<case>
4

31) noun group translation example

viele Museen Hochschulen und Theater	<case>
wiele muzeów szkółwyższych i teatrów	1
wielu muzeów szkółwyższych i teatrów	2
wielu muzeom szkołom wyższym i teatrom	3
wiele muzeów szkółwyższych i teatrów	4
wieloma muzeami szkołami wyższymi i teatrami	5
wielu muzeach szkołach wyższych i teatrach	6

<person>	<number>	<gender>
3	2	3

The effect of the work of the proposed example-based machine translation system is Polish translation of the original German text:

Berlin zostałwspomniany w dokumentach po raz pierwszy w roku 1244. Berlin byłw tym czasie bardzo małą wsią. Stopniowo jednakże Berlin stawałsię coraz większy. Berlin zostałstolicą Prus w roku 1740. Berlin uzyskałcentralne znaczenie dla Niemiec w roku 1871. Berlin zostałstolicą Rzeszy. Berlin ma jako miasto kultury sławę międzynarodową. Tutaj znajduje się wiele muzeów szkółwyższych i teatrów.

It must be stressed that the Polish translation is both correct from the grammatical point of view and it is an exact translation of the original German text. Moreover, the obtained Polish text seems to be natural, and thus very similar to the one produced by a human translator. These facts point out that the modified example-based machine translation technique proposed by this author is headed in the right direction.

The above example of machine translation results between German and Polish is of course not the only one that can be obtained by the system developed by this author. This author has gathered quite a big database of translation example by the medium of which also other simple German texts can be translated into Polish. The machine translation system is still under development and new items are systematically added to the translation examples database.

8. Final Conclusions

The high-quality machine translation system for unrestricted text has always been an unachievable goal for the computer scientists working in the field of automatic translation between human languages. And maybe, because of the reasons of fundamental nature (the lack of possibility of constructing an algorithm equivalent to the creativeness of the human mind) human translators will never be eliminated by computers totally, and high-quality machine translation for unrestricted text will forever remain the Holy Grail of scientific research (Mitamura *et al.*, 1998). But, by using various machine translation techniques we can of course try to approach as close as possible to this unattainable goal (Loukachevitch and Dobrov, 2000). Quite recently the example-based machine translation technique has emerged as a very serious and tempting alternative to the existing systems that are mainly based on the knowledge developed in the field of computational linguistics.

In the paper the implementation of example-based machine translation technique in the system, which translates from German into Polish is proposed. In order to use the example-based machine translation technique for the Polish language, which posses very specific grammatical features, so different from the West-European languages, this author proposed a thorough modification of this technique that allows to take into account the flexion nature of the Polish language.

The results obtained so far are very promising and show that the usage of example-based machine translation technique for the Polish and German language pair is a step made in the right direction. But, we cannot forget that the final success depends strongly on the dimension of the database of translation examples. The effective constructing of such database requires a lot of work and time. In fact, it is a task for a quite big teem of trained linguists and computer scientist, who basing only on the great bilingual corpus would be able to extract all necessary and most frequently used translation examples.

Last but not least, this author would like to mention that according with his knowledge the proposed German-to-Polish machine translation system is the first system of this kind that has ever been built, thus the results obtained by this author have a totally pioneer character.

References

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

K.L., Baker, K.L., A.M. Franz, P.W. Jordan, T. Mitamura, E.H. Nyberg (1998). *Coping with Ambiguity in a Large-Scale Machine Translation System*. Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA.

Bandyopadhyay, S. (2000). State and role of machine translation in India. *Machine Translation Review*, **11**, 1–3.

Blekhman, M., B. Pevzner (2000). First steps of language engineering in the USSR: the 50s tthrough 70s. *Machine Translation Review*, **11**, 5–7.

Canals, R., A. Esteve, A. Garrido, M.I. Guardiola, A. Iturraspe–Bellver, S. Montserrat (2000). InterNOSTRUM: a Spanish–Catalan machine translation system. *Machine Translation Review*, **11**, 21–25.

Carbonell, J.G., T. Mitamura, E.H. Nyberg (1998). *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, . . .)*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA.

Fukutomi, O. (2000). Report on commercial machine translation in a manufacturing domain. *Machine Translation Review*, **11**, 16–25.

Loukachevitch, N.V., B.V. Dobrov (2000). Thesaurus-based structural thematic summary in multilingual information systems. *Machine Translation Review*, **11**, 10–20.

Majewicz, A.F. (1989). *The Languages of the World and their Classifying*, PWN, Warsaw, Poland (in Polish).

Melby, A. (1999). Machine translation and philosophy of language. *Machine Translation Review*, **9**, 6–17.

Mitamura, T. (1998). *Controlled Languages for Multilingual Machine Translation*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA.

Mitamura, T., E.H. Nyberg, J. Carbonell (1998). *An Efficient Interlingua Translation System for Multi-Lingual Document Production*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA.

Murphy, D. (2000). Keeping translation technology under control. *Machine Translation Review*, **11**, 7–10.

Ney, H., S. Nießen, F.J. Och, H. Sawaf, C. Tillmann, S. Vogel (2000). Algorithms for statistical translation of spoken language. IEEE *Transactions on Speech and Audio Processing*, **8**(1), 24–36.

Nyberg, E., T. Mitamura, J. Carbonell (1998). *The KANT Machine Translation System: From R&D to Initial Deployment*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA.

Penrose, R. (1995). *New Caesar's Mind*, PWN, Warsaw, Poland (in Polish).
Rico, C. *From Novelty to Ubiquity: Computers and Translation at the Close of Industrial Age*, `http://www.accurapid.com/journal/15mt2.htm`
Zue, V.W., J.R. Glass (2000). Conversational interfaces: advances and challenges. In *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1166–1180.
Waibel, A., P. Geutner, L.M. Tomokiyo, T. Schultz, M. Woszczyna (2000). Multilinguality in speech and spoken language systems. In *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313.
Whitelock, P., K. Kilby (1995). *Linguistic and Computational Techniques in Machine Translation System Design*, UCL Press, London, GB.

M. Gajer was born in Cracow (Poland) 1971. In 1996 he obtained a Msc. degree in the field of electronics, and in 2000 he obtained his PhD degree in the field of computer science. Now he is with the Automatic Control Department at the Technical University in Cracow. His research interest field is artifficial intelegence, especially natural language processing, machine translation and neural networks.

Pavyzdžiais pagrįsto automatinio vertimo metodo realizacija vokiečių-lenkų automatinio vertimo sistemai

Mirosław GAJER

Automatinis geros kokybės vertimas iš vienos kalbos į kitą ilgą laiką buvo nepasiekiama mokslininkų, dirbančių šioje patrauklioje tarpdalykinėje kompiuterių taikymo srityje. Neseniai sukurtas pavyzdžiais pagrįstas automatinio vertimo metodas turėtų tapti rimti iki šiol buvusių automatinio vertimo metodų alternatyva. Šiame straipsnyje pasiūlytas pavyzdžiais pagrįsto automatinio vertimo metodo panaudojimas neriboto teksto vertimo iš vokiečių į lenkų kalbą sistemos sukūrimui. Pradiniai pasiūlytos sistemos taikymo rezultatai atrodo daug žadantys ir yra žingsnis teisinga kryptimi link visiškai automatinio geros kokybės neriboto teksto vertimo iš vokiečių į lenkų kalbą sistemos.