Control and Cybernetics

vol. 34 (2004) No. 2

The implementation of the example-based machine translation technique for Norwegian-to-Polish automatic translation system

by

Mirosław Gajer

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

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 application of computers. The quite recently developed 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 a system, which would be able to translate an unrestricted Norwegian 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 development of the proposed system seem to be very promising and appear to be a step made in the right direction towards a fully-automatic high quality Norwegian-into-Polish machine translation system aimed for an unrestricted text.

Keywords: natural language processing, computational linguistics, machine translation.

1. Introduction

Soon after the invention of the first digital computer in 1946 some non-numerical applications were proposed, among which there was also machine translation. Machine translation is a science (or maybe still an art) that delivers the knowledge about how to program the computers, to make them able to translate between human languages, for example, between Danish and Polish. It may be

of computer itself (Blekhman, Pevzner, 2000). In 1949 an American scientist Warren Weaver sent the memorandum to The Rockefeller Foundation, 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).

In the paper this author presents the machine translation technique developed by himself, which is a modified example-based machine translation technique, aimed especially on Polish as a target language. The paper is organized as follows. In Section 2 a short outline of the history of development of machine translation systems is given. In Section 3 it is explained why automation of translation is such a difficult task, and why despite of more than 50 years of scientific effort a fully automatic high-quality machine translation system is still rather a dream than reality. In Section 4 the way in which example-based machine translation technique must be modified when considering Polish as a target language is presented. In Section 5 the results obtained during the implementation of the developed machine translation technique for Norwegian as a source language are presented. The paper is concluded in Section 6.

2. An outline of machine translation history

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 (Perez, no date). 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 to 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 (Canals, et al., 2000), and India (Bandyopadhyay, 2000), but after so many

human languages for unrestricted text is still a long-term scientific dream of enormous political, social, and scientific importance (Mitamura, 1999). 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).

3. Automatic translation - a dream or reality?

Despite more than 50 years of intensive scientific research the automation of translation is still far away from its fully satisfactory solution. 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 (Majewicz, 1989).

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, 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 is clear that the personal pronouns system of Arabic is much more complicated than the one of Hungarian. It is caused by the fact that Hungarian language does not know such invention as grammatical gender of 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 full automation of this process is still a pure science-fiction.

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 complex system. If we want to translate, for example, from Chinese to Croatian, it is a hard work to find in Croatian the equivalents of Chinese words that preserve their original meanings. While doing this a human translator has to cope with enormous number of lexical holes, that is, words that do not have their equivalents in the other language, and as such can be translated only by the medium of a long description that clears up their semantics.

The way the different languages arrange lexical units with respect to the real world objects and abstract notions 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 numbered from 1 to 6. Further, we have two different natural languages: language A and language B.

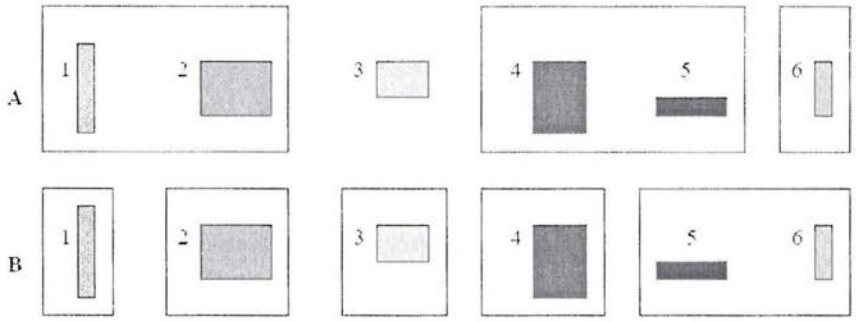

Figure 1. Illustration of the way in which different languages divide reality into lexical items.

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 object 3 does not 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,

one lexical entity.

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

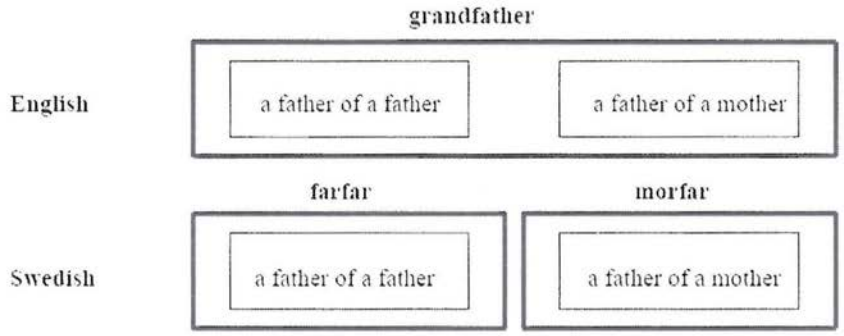

Figure 2. English word *grandfather* versus Swedish words *farfar* and *morfar*.

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 the French word *fleuve*, and in the second case the correct choice is French word *rivière*. This situation is illustrated in Fig. 3.

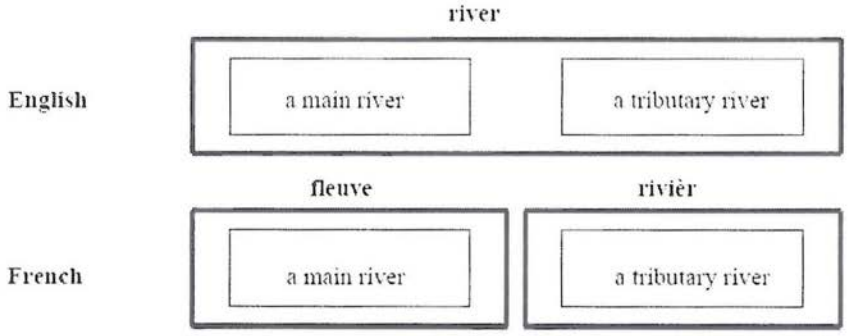

Figure 3. English word *river* versus French words *fleuve* and *rivière*.

But the perhaps most serious problem, which the computer has to cope with in machine translation is the ambiguity of human languages (Baker et al., 1999; Net et al., 2000). 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 do not 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 different. The problem is that this ambiguity cannot be preserved during 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 woman.

When analyzing this phrase we do not know whether it is equivalent to:

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 (Whitelook, Kilby, 1995). A good example is the English sentence:

She threw the vase at the window and it broke.

This sentence is ambiguous because we do not know what is broken, *a window* or *a vase*? If we wished 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:

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 is 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 of ten different words, and let each of these words have 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 in this way we obtain a correct translation of a whole document built from many such sentences is equal in practice to zero. Moreover, no efficient algorithm that allows for solving this problem is known, and lexical ambiguity can be found in abundance in any human language - some examples of possible Polish translations of lexically ambiguous words taken from several languages of the world are listed below.

Polish equivalents of the French word **perle** are: 1. perła, 2. paciorek, 3. kapsułka
Polish equivalents of the Spanish word **fondo** are: 1. dno, 2. głębia, 3. tło
Polish equivalents of the Italian word **stufa** are: 1. piec, 2. cieplarnia
Polish equivalents of the German word **Absatz** are: 1. ustęp, 2. obcas, 3. osad, 4. złoże, 5. osadzenie, 6. zbyt
Polish equivalents of the English 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 the Dutch word **boodschap** are: 1. poselstwo, 2. polecenie, 3. wiadomość, 4. zakupy
Polish equivalents of the Swedish word **tomten** are: 1. parcela, 2. plac, 3. krasnoludek
Polish equivalents of the Norwegian word **hytte** are: 1. chata, 2. szałas, 3. buda, 4. huta, 5. kabina
Polish equivalents of the Danish word **løber** are: 1. biegacz, 2. dywanik
Polish equivalents of the Finnish word **kanta** are: 1. podstawa, 2. obcas, 3. stanowisko, 4. baza
Polish equivalents of the Greek word $\sigma\kappa o\pi o\varsigma$ (skopos) are: 1. zamiar, 2. melodia, 3. wartownik
Polish equivalents of the Arabic word **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:

This nominal group can be understood as:

manufacturer of toys for adults

or

an adult manufacturer of toys.

Another kind of difficulties making automation of the translation process so hard are all idiomatic phrases. The problem is that these idiomatic phrases can be also interpreted literally. 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 literally it means that:

The town is beautiful.

The question is, 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.

Taking under consideration all the 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 *The Emperor's New 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, Mitamura, Carbonell, 1999; Mitamura, 1999)?

The philosopher Alan 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 field 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). The arguments given by Roger Penrose are very strong

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 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 translation having 99% of 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 translation 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. (Carbonell, Mitamura, Nyberg, 1999; Ney et al., 1999; Canals et al., 2000; Loukachevich, Dobrov, 2000). Among these approaches example-based machine translation is becoming a serious alternative paradigm (Brown, 2001; Carl, 2001; Menezes, Richardson, 2001; Way, 2001; Somers, 2001; McTait, 2001; Turcato, Popowich, 2001).

The idea of example-based machine translation is very simple. Let there be given any bilingual text. For example Polish and English:

"Upon my return to the United States a few months ago, ***after the extraordinary series of adventures*** *in the South Seas and elsewhere, of which an account is given in the following pages, accident threw me into the society of several gentlemen in Richmond, Va., who felt deep interest in all maters relating to the region and who were constantly urging it upon me, as a duty, to give my narrative to the public."*

"Kilka miesięcy temu, gdy wróciłem do Stanów Zjednoczonych ***po wielu nadzwyczajnych przygodach*** *na morzach południowych i w innych okolicach, o czym będę mówił dalej, przypadkowo znalazłem się w Richmond, stan Wirginia, w towarzystwie kilku dżentelmenów, których ogromnie zainteresowały wszystkie szczegóły zwiedzanych przeze mnie okolic i którzy ustawicznie wmawiali mi, że obowiązkowo powinienem opublikować moje wspomnienia."*

The upper text is the first sentence by the story of Edgar Allan Poe *The Narrative of Arthur Gordon Pym of Nantucket* and the lower text is its translation made by a professional human translator.

In the upper text the phrase "after the extraordinary series of adventures" was translated by the human translator as "po wielu nadzwyczajnych przygodach". The main idea of the example-based translation is that if such an English phrase appears once again in any English text being automatically translated, and machine translation systems substitutes this English phrase with its Polish equivalent "po wielu nadzwyczajnych przygodach", we obtain once again a correct translation (Somers, 2001). This is only a supposition, not a certainty, but example-based machine translation technique has proved many times to the a serious alternative to the other existing methods Carl (2001).

Up till now there exist known cases in which the effort put into the develop-

ment of a machine translation system based on the example-based technique has lead to a construction of a high quality machine translation systems that were used in practice. 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 system 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 (Perez, 2000).

4. Modifying example-based machine translation technique for Polish as a target language

Polish language, which belongs to the group of 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 would not 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:

- In most human languages one can distinguish the first, the second, and the third grammatical person.
- In most human languages one can distinguish such elements of 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 Polish language the grammatical 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>, <gender>. During the process of translation the values of these attributes

are set, respectively, to the grammatical rules of the Polish language, so that the subject and the verb agree in the 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:

- Noun group translation examples. These translation examples play the role of the subject or the object of the sentence, which is to be translated.
- Verb group translation examples. These translation examples play the role of the verb of the sentence, which is to be translated.
- Non-flexion type translation examples.

5. Implementing the system for Norwegian as a source language

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

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

The objective of the proposed system is to translate to Polish the following Norwegian text:

Det er fire universiteter i Norge. Universitetet i Tromso er i Nord-Norge, Universitetet i Bergen er på Vestlandet, Universitetet i Trondheim er i Trøndelag, og Universitetet i Oslo er pa Østlandet. Universitetet i Oslo ligger egentlig pa to steder i Oslo. Det gamle universitetet fra 1811 ligger need i sentrum og det nye universitetet ligger på Blindern. Blindern er et lite stykke utenfor sentrum. Nesten alle studenter er nå på Blindern. Det nye universitetet på Blindern er et stort og moderne sted. Det er nesten en liten by for studentene. Det er en enorm kafeteria med en stor grill og en liten restaurant. Kafeterian som heter Frederike, er et populart sted. Studentene finner nesten alt de trenger på Blindern. Men de kann egentlig ikke bo der. Det er bare en liten gruppe som bor på Blindern Studenterhjem. De fleste bor på Kringsjå og på Sogn, et stykke fra Universitetet.

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

<case> = 1;

1) non-flexion type translation example

det er fire universiteter
są cztery uniwersytety

2) non-flexion type translation example

i Norge
w Norwegii

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

<case> = 1;

3) noun group translation example

Universitetet i Tromsø	<case>
Uniwersytet w Tromsø	1
Uniwersytetu w Tromsø	2
Uniwersytetowi w Tromsø	3
Uniwersytet w Tromsø	4
Uniwersytetem w Tromsø	5
Uniwersytecie w Tromsø	6
<number> = 1; <gender> = 1;	

4) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

5) non-flexion type translation example

i Nord-Norge
w Norwegii Północnej

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

<case> = 1;

6) noun group translation example

Universitetet i Bergen	<case>
Uniwersytet w Bergen	1
Uniwersytetu w Bergen	2
Uniwersytetowi w Bergen	3
Uniwersytet w Bergen	4
Uniwersytetem w Bergen	5
Uniwersytecie w Bergen	6
<number> = 1; <gender> = 1;	

7) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

8) non-flexion type translation example

på Vestlandet
w Okręgu Zachodnim

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

<case> = 1;

9) noun group translation example

Universitetet i Trondheim	<case>
Uniwersytet w Trondheim	1
Uniwersytetu w Trondheim	2
Uniwersytetowi w Trondheim	3
Uniwersytet w Trondheim	4
Uniwersytetem w Trondheim	5
Uniwersytecie w Trondheim	6
<number> = 1; <gender> = 1;	

10) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

11) non-flexion type translation example

i Trondelag
w Okręgu Trondelag

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

<case> = 1;

12) noun group translation example

Universitetet i Oslo	<case>
Uniwersytet w Oslo	1
Uniwersytetu w Oslo	2
Uniwersytetowi w Oslo	3
Uniwersytet w Oslo	4
Uniwersytetem w Oslo	5
Uniwersytecie w Oslo	6
<number> = 1; <gender> = 1;	

13) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

14) non-flexion type translation example

pa Ostlandet
w Okręgu Wschodnim

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

<case> = 1;

15) noun group translation example

Universitetet i Oslo	<case>
Uniwersytet w Oslo	1
Uniwersytetu w Oslo	2
Uniwersytetowi w Oslo	3
Uniwersytet w Oslo	4
Uniwersytetem w Oslo	5
Uniwersytecie w Oslo	6
<number> = 1; <gender> = 1;	

16) verb group translation example

ligger	<number>	<gender>
leży	1	1
leży	1	2
leży	1	3
leżą	2	1
leżą	2	2
leżą	2	3
<case> = 5;		

17) non-flexion type translation example

egentligen
wąściwie

18) non-flexion type translation example

på to steder
w dwóch miastach

19) non-flexion type translation example

i Oslo
w Oslo

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

<case> = 1;

20) noun group translation example

det gamle universitetet	<case>
stary uniwersytet	1
starego uniwersytetu	2
staremu uniwersytetowi	3
stary uniwersytet	4
starym uniwersytetem	5
starym uniwersytecie	6
<number> = 1;	

21) non-flexion type translation example

fra 1811
z roku 1811

22) verb group translation example

ligger	<number>	<gender>
leży	1	1
leży	1	2
leży	1	3
leżą	2	1
leżą	2	2
leżą	2	3
<case> = 5;		

23) non-flexion type translation example

need i sentrum
w centrum

24) non-flexion type translation example

og
i

25) noun group translation example

det nye universitetet	<case>
nowy uniwersytet	1
nowego uniwersytetu	2
nowemu uniwersytetowi	3
nowy uniwersytet	4
nowym uniwersytetem	5
nowym uniwersytecie	6
<number> = 1; <gender> = 1;	

26) verb group translation example

ligger	<number>	<gender>
leży	1	1
leży	1	2
leży	1	3
leżą	2	1
leżą	2	2
leżą	2	3
<case> = 5;		

27) non-flexion type translation example

på Blindern

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

<case> = 1;

28) noun group translation example

Blindern	<case>
dzielnica Blindern	1
dzielnicy blindern	2
dzielnicy Blindern	3
dzielnicą Blindern	4
dzielnicą Blindern	5
dzielnicy Blindern	6
<number> = 1; <gender> = 2;	

29) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

30) noun group translation example

et lite stykke utenfor sentrum	<case>
mała dzielnica leżąca poza centrum	1
małej dzielnicy leżącej poza centrum	2
małej dzielnicy leżącej poza centrum	3
małą dzielnicą leżącą poza centrum	4
małą dzielnicą leżącą poza centrum	5
małej dzielnicy leżącej poza centrum	6
<number> = 1; <gender> = 2;	

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

<case> = 1;

31) non-flexion type translation example

nesten
prawie

32) noun group translation example

alle studenter	<case>
wszyscy studenci	1
wszystkich studentów	2
wszystkim studentom	3
wszystkich studentów	4
wszystkimi studentami	5
wszystkich studentach	6
<number> = 2; <gender> = 1;	

33) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

34) non-flexion type translation example

na
teraz

35) non-flexion type translation example

på Blindern
w dzielnicy Blindern

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

<case> = 1;

36) noun group translation example

det nye universitetet	<case>
nowy uniwersytet	1
nowego uniwersytetu	2
nowemu uniwersytetowi	3
nowy uniwersytet	4
nowym uniwersytetem	5
nowym uniwersytecie	6
<number> = 1; <gender> = 1;	

37) non-flexion type translation example

på Blindern
w dzielnicy Blindern

38) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

39) noun group translation example

et stort og moderne sted	<case>
duże i nowoczesne miejsce	1
dużego i nowoczesnego miejsca	2
dużemu i nowoczesnemu miejscu	3
duże i nowoczesne miejsce	4
dużym i nowoczesnym miejscem	5
dużym i nowoczesnym miejscu	6
<number> = 1; <gender> = 3;	

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

<case> = 1;

40) non-flexion type translation example

det er
jest tam
<case> = 1;

41) non-flexion type translation example

nesten
prawie

42) noun group translation example

en liten	<case>
małe miasto	1
małego miasta	2
małemu miastu	3
małe miasto	4
małym miastem	5
małym mieście	6
<number> = 1; <gender> = 1;	

43) non-flexion type translation example

for studentene
dla studentów

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

<case> = 1;

44) non-flexion type translation example

det er
jest tam
<case> = 1;

45) noun group translation example

en enorm kafeteria	<case>
olbrzymia kawiarnia	1
olbrzymiej kawiarni	2
olbrzymiej kawiarni	3
olbrzymią kawiarnią	4
olbrzymią kawiarnią	5
olbrzymiej kawiarni	6
<number> = 1; <gender> = 2;	

46) non-flexion type translation example

med en stor grill
z dużym grillem

47) non-flexion type translation example

og
i

48) noun group translation example

en liten restaurant	<case>
mała restauracja	1
małej restauracji	2
małej restauracji	3
małą restauracją	4
małą restauracją	5
małej restauracji	6
<number> = 1; <gender> = 2;	

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

<case> = 1;

49) noun group translation example

kafeterian	<case>
kawiarnia	1
kawiarni	2
kawiarni	3
kawiarnią	4
kawiarnią	5
kawiarni	6
<number> = 1; <gender> = 2;	

50) verb group translation example

som heter Frederike	<case>
który nazywa się Fryderyka	1
która nazywa się Fryderyka	2
które nazywa się Fryderyka	3
którzy nazywają się Fryderyka	4
które nazywają się Fryderyka	5
które nazywają się Fryderyka	6

51) verb group translation example

er	<number>	<gender>
jest	1	1
jest	1	2
jest	1	3
są	2	1
są	2	2
są	2	3
<case> = 5;		

52) noun group translation example

et populart sted	<case>
popularne miejsce	1
popularnego miejsca	2
popularnemu miejscu	3
popularne miejsce	4
popularnym miejscem	5
popularnym miejscu	6
<number> = 1; <gender> = 3;	

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

<przypadek> = 1;

53) noun group translation example

studentene	<case>
studenci	1
studentów	2
studentom	3
studentów	4
studentami	5
studentach	6
<number> = 2; <gender> = 1;	

54) verb group translation example

finner	<number>	<gender>
znajduje	1	1
znajduje	1	2
znajduje	1	3
znajdują	2	1
znajdują	2	2
znajdują	2	3
<case> = 5;		

55) non-flexion type translation example

nesten
prawie

56) noun group translation example

alt de trenger	<case>
wszystko czego potrzebują	1
wszystkiego czego potrzebują	2
wszystkiemu czego potrzebują	3
wszystko czego potrzebują	4
wszystkim czego potrzebują	5
wszystkim czego potrzebują	6
<number> = 1; <gender> = 3;	

57) non-flexion type translation example

på Blindern
w dzielnicy Blindern

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

<case> = 1;

58) non-flexion type translation example

men
ale

59) non-flexion type translation example

de
oni
<number> = 2; <gender> = 1;

60) verb group translation example

kann egentlig ikke bo	<case>	<gender>
nie może właściwie mieszkać	1	1
nie może właściwie mieszkać	1	2
nie może właściwie mieszkać	1	3
nie mogą właściwie mieszkać	2	1
nie mogą właściwie mieszkać	2	2
nie mogą właściwie mieszkać	2	3

61) non-flexion type translation example

der
tam

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

<case> = 1;

62) non-flexion type translation example

det er
jest tam
<case> = 1;

63) non-flexion type translation example

bare
zaledwie

64) noun group translation example

en liten gruppe	<case>
mała grupa	1
małej grupy	2
małej grupie	3
małą grupą	4
mała grupą	5
małej grupie	6
<number> = 1; <gender> = 2;	

65) verb group translation example

som	<case>	<gender>
któr	1	1
która	1	2
które	1	3
którzy	2	1
które	2	2
które	2	3

66) verb group translation example

bor på	<case>	<gender>
mieszka w	1	1
mieszka w	1	2
mieszka w	1	3
mieszkają w	2	1
mieszkają w	2	2
mieszkają w	2	3
<number> = 6;		

67) noun group translation example

Blindern Studenterhjem	<case>
dom studencki w dzielnicy Blindern	1
domu studenckiego w dzielnicy Blindern	2
domowi studenckiemu w dzielnicy Blindern	3
dom studencki w dzielnicy Blindern	4
domem studenckim w dzielnicy Blindern	5
domu studenckim w dzielnicy Blindern	6
<number> = 1; <gender> = 1;	

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

<case> = 1;

68) non-flexion type translation example

de
oni
<number> = 2; <gender> = 1;

69) non-flexion type translation example

fleste
najczęściej

70) verb group translation example

bor på	<case>	<gender>
mieszka w	1	1
mieszka w	1	2
mieszka w	1	3
mieszkają w	2	1
mieszkają w	2	2
mieszkają w	2	3
<number> = 6;		

71) noun group translation example

Kringsja	<case>
dzielnica Kringsja	1
dzielnicy Kringsja	2
dzielnicy Kringsja	3
dzielnicą Kringsja	4
dzielnicą Kringsja	5
dzielnicy Kringsja	6
<number> = 1; <gender> = 2;	

72) non-flexion type translation example

og
i

73) non-flexion type translation example

på Sogn
w dzielnicy Sogn

74) noun group translation example

et stykke fra Universitetet	<case>
część uniwersytetu	1
części uniwersytetu	2
części uniwersytetu	3
część uniwersytetu	4
częścią uniwersytetu	5
części uniwersytetu	6
<number> = 1; <gender> = 2;	

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

Są cztery uniwersytety w Norwegii. Uniwersytet w Tromso jest w Norwegii Północnej. Uniwersytet w Bergen jest w Okręgu Zachodnim. Uniwersytet w Trondheim jest w Okręgu Trondelag. Uniwersytet w Oslo jest w Okręgu Wschodnim. Uniwersytet w Oslo leży właściwie w dwóch miastach w Oslo. Stary uniwersytet z roku 1811 leży w centrum i nowy uniwersytet leży w dzielnicy Blindern. Dzielnica Blindern jest małą dzielnicą leżącą poza centrum. Prawie wszyscy studenci są teraz w dzielnicy Blindern. Nowy uniwersytet w dzielnicy Blindern jest dużym i nowoczesnym miejscem. Jest tam prawie małe miasto dla studentów. Jest tam olbrzymia kawiarnia z dużym grillem i mała restauracja. Kawiarnia która nazywa się Fryderyka jest popularnym miejscem. Studenci znajdują prawie wszystko czego potrzebują w dzielnicy Blindern ale oni nie mogą właściwie mieszkać tam. Jest tam zaledwie mała grupa która mieszka w domu studenckim w dzielnicy Blindern. Oni najczęściej mieszkają w dzielnicy Kringsja i w dzielnicy Sogn części uniwersytetu.

It must be stressed that the Polish translation is both correct from the grammatical point of view and it is also an exact translation of the original Norwegian 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. Of course, the obtained Polish translation is not perfect, but the question is if the perfect translation does really exist? If we compare the effects of work of different human translators that have

as better, while the other ones as worse. But in many cases the assessment of translation quality is a very subjective matter, and even some experts may differ in their opinions. Taking the above into account the conclusion is that perfect translation does not exist and it is only the indication of direction, at which the process of translation should be aimed. The modified example-based machine translation technique proposed by this author is a method that allows to approach quite close to this aim of perfect translation which can be never attained either by a human or a computer.

6. Final remarks

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, Nyberg, Carbonell, 1999). But, by using various machine translation techniques we can of course try to approach as close as possible this unattainable goal (Loukachevich, 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 an implementation of the example-based machine translation technique in the system, which translates from Norwegian to Polish, is proposed. In order to use the example-based machine translation technique for the Polish language, which possesses very specific grammatical features, this author proposed a 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 Norwegian language pair is a step made in the right direction. But, one cannot forget that the final success depends strongly on the dimension of the database of translation examples. The effective construction of such a database requires a lot of work and time. In fact, it is a task for a quite big team of trained linguists and computer scientists, who, basing only on a great bilingual corpus would be able to extract all the necessary and most frequently used translation examples.

The results obtained so far by this author are very promising. Using the example-based machine translation technique this author has translated into Polish a vast corpus of foreign texts. The statistics concerning the translation examples length (Fig. 4) seems to be very interesting.

In order to translate into Polish the corpus of foreign texts totally 2737

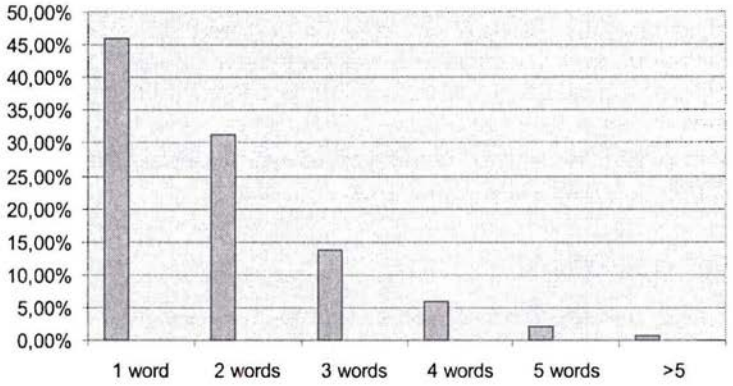

Figure 4. Statistics of the translation examples lengths.

1258 one-word examples, 856 two-word examples, 377 three-word examples, 156 four-word examples, 59 five-word examples, and only 22 translation examples that were longer than five words. The translation examples of the length equal one, two or three words constitute together 91.01% of all the translation examples. Hence, the average length of translation examples is not big. Moreover, almost everything can be translated by the use of translation examples that are not longer than three or four words. From this point of view one can be optimistic about the possibility of gathering a database of translation examples that would be sufficient for performing translation of a sufficiently good quality.

References

ARNOLD, D., BALKAN, L., MEIJER, S., HUMPHREYS, R.L. and SADLER, L. (1994) *Machine Translation.* An Introductory Guide, NCC Blackwell, London.

BAKER, K.L., FRANZ, A.M., JORDAN, P.W., MITAMURA T. and NYBERG, E.H. (1999) *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. and PEVZNER, B. (2000) First Steps of Language Engineering in the USSR, The 50s Through 70s. *Machine Translation Review* **11**, 5-7.

BROWN, R.D. (2001) Transfer-Rule Induction for Example-Based Translation, *MT Summit VIII.* Santiago de Compostela, http://www.eamt.org/summitVIII/workshop-papers.html.

CANALS, R., ESTEVE, A., GARRIDO, A., GUARDIOLA, M.I., ITURRASPE-BELLVER, A. and MONTSERRAT S. (2000) InterNOSTRUM: A Spanish-Catalan Machine Translation System. *Machine Translation Review* **11**, 21-25.

CARBONELL, J.G., MITAMURA, T. and NYBERG, E.H. (1999) *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, ...).* Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA.

CARL, M. (2001) Inducing Translation Grammars for Aligned Text, *MT Summit VIII.* Santiago de Compostela, http://www.eamt.org/summitVIII/workshop-papers.html.

FUKUTOMI, O. (2000) Report on Commercial Machine Translation in a Manufacturing Domain, *Machine Translation Review* **11**, 16-25.

LOUKACHEVITCH, N.V. and DOBROV, B.V. (2000) Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems. *Machine Translation Review* **11**, 10-20.

MAJEWICZ, A.F. (1989) *Languages of The World and Their Classifying.* PWN. Warsaw, Poland.

MCTAIT, K. (2001) Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns. *MT Summit VIII.* Santiago de Compostela, http://www.eamt.org/summitVIII/workshop-papers.html.

MELBY, A. (1999) Machine Translation and Philosophy of Language. *Machine Translation Review* **9**, 6-17.

MENEZES, A. and RICHARDSON, S. (2001) A Best-First Alignment Algorithm for Automatic Extraction of Transfer Mapping from Bilingual Corpora. *MT Summit VIII.* Santiago de Compostela, http://www.eamt.org/summitVIII/workshop-papers.html.

MITAMURA, T. (1999) *Controlled Languages for Multilingual Machine Translation.* Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA.

MITAMURA, T., NYBERG, E.H. and CARBONELL, J. (1999) *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., NIESSEN, S., OCH, F.J., SAWAF, H., TILLMANN, C. and VOGEL S. (2000) Algorithms for Statistical Translation of Spoken Language *IEEE Transactions on Speech and Audio Processing* **8** (1), 24-36.

NYBERG, E., MITAMURA, T. and CARBONELL, J. (1999)The KANT Machine Translation System: From R&D to Initial Deployment. Center for Ma-

PENROSE, R. (1995) *The Emperor's New Mind.* PWN, Warsaw, Poland.

PEREZ, C.R. (2000) *From Novelty to Ubiquity: Computers and Translation at the Close of Industrial Age.* http://www.accurapid.com/journal/15mt2.htm.

SOMERS, H. (2001) EBMT Seen as Case-Based Reasoning. *MT Summit VIII.* Santiago de Compostela, http://www.eamt.org/summitVIII/workshop-papers.html.

TURCATO, D. and POPOWICH, F. (2001) What is Example-Based Machine Translation? *MT Summit VIII.* Santiago de Compostela, http://www.eamt.org/summitVIII/workshop-papers.html.

WAIBEL, A., GEUTNER, P., TOMOKIYO, L.M., SCHULTZ, T., and WOSZCZYNA, M. (2000) Multilinguality in Speech and Spoken Language Systems. *Proceedings of the IEEE* **88** (8), 1297-1313.

WAY, A. (2001) Translating with Examples. *MT Summit VIII.* Santiago de Compostela, http://www.eamt.org/summitVIII/workshop-papers.html.

WHITELOCK, P. and KILBY, K. (1995) *Linguistic and Computational Techniques in Machine Translation System Design.* UCL Press, London, GB.

ZUE, V.W. and GLASS, J.R. (2000) Conversational Interfaces: Advances and Challenges. *Proceedings of the IEEE* **88** (8), 1166-1180.