SPECIALIZED FULLY AUTOMATIC MACHINE TRANSLATION SYSTEM DELIVERING HIGH QUALITY OF TRANSLATED TEXTS

MIROSŁAW GAJER

University of Science and Technology, Department of Control Science,
Al. Mickiewicza 30, 30-059 Cracow, Poland
mgajer@ia.agh.edu.pl

(Received 18 September 2009; revised manuscript received 30 September 2009)

Abstract: The paper concerns machine translation systems that form a discipline of computer science and are aimed at writing computer programs that are able to translate text between natural languages. In the paper the author argues that it is not possible to build a machine translation system that would be able to translate any kind of documents with a sufficiently high quality. Instead, the author proposes a specialized machine translation system the aim of which is to translate financial reports concerning the global currency exchange market – forex. For the purpose of building the above mentioned system, the author has proposed his own machine translation method of translation patterns. The translation patterns allow transferring the translation process from the level of single words to the level of words chunks. The translation patterns play a very important role in the case of such an inflectional language as Polish because they make it possible to choose the correct form of Polish translation of foreign phrases depending whether they perform the verb or object function in the sentence. The high quality of the specialized machine translation system developed by the author was proved with many experiments the results of which are demonstrated in the paper. The quality of translation is so high that the Polish translations of English reports from the global currency exchange market can be published on Web pages without any additional changes. Thus, it is possible to totally eliminate the human translator from the process of translation of texts which are highly stereotypical and oriented to a selected and narrow domain.

Keywords: natural language processing, machine translation, translation patterns

1. Introduction

The aim of research work within the discipline of machine translation systems is to develop computer programs that would be able to translate texts written in one natural language into another natural language [1]. The origins of machine translation research work go back into the early 50s. The first machine translation research group was established in 1951 and the first public demonstration of the operating machine translation system took place in 1954 [2]. Despite the many years of scientific research a fully-automatic high-quality machine translation system still seems to be an unattainable goal for intensive scientific research.

From the perspective of the current state of the art in machine translation systems, it does not seem possible any longer that the idea to eliminate the human translator totally and to replace him by a computer program could be implemented in the future. Moreover, machine translation software is often seen only as an aid for a human translator or it can be helpful for people that do not know a foreign language at all and want to get the gist of a document, possibly to decide, whether it is worth engaging a professional human translator to translate it precisely. However, it is probably possible in some special cases to replace a human translator by a computer, but it seems to be possible only in the case of specialized machine translation systems. The aim of such specialized machine translation systems is not to translate everything, *i.e.* any kind of a written document, but only some special sorts of documents. In this way the scope of documents to be translated by the computer must be restricted to a precisely defined discipline of human activity. Moreover, it is only a strict restriction of the subject matter of the translated documents that can lead to a fully-automatic high-quality machine translation system that can eliminate the human translator, but only in the domain of this special area of translation [3].

The topic of the paper is a specialized fully-automatic machine translation system the aim of which is to eliminate human translators in translation of documents from the forecasts for the foreign currency exchange market (forex). A specialized machine translation system which is described in the paper is based on the Pattern-Based Machine Translation (PBMT) method that was developed by the author. The PBMT method takes advantage of both structural and cognitive approaches to the process of translation because it uses syntactic structures to build grammatically correct sentences and it translates texts on a higher level than the level of single words [4].

The paper is organized as follows. The first section is the introduction to the subject of the paper. The second section presents the Pattern-Based Machine Translation method which was developed by the author. In section three a report of the construction of a specialized machine translation system that was developed by the author is described. Section four concludes the paper.

2. Pattern-Based Machine Translation

In the case of classical Rule-Based Machine Translation (RBMT) systems the process of translation is performed on the level of single words. In such systems it must be determined for each word what role it plays in the sentence, thus, the grammar tree is built. In the next stage a syntactical transfer takes place during which syntactical structures of the source language are converted into those typical for the target language. In that process words of the source language are substituted with their equivalents in the target language using a bilingual dictionary.

In the 90s a new approach to machine translation appeared which is currently known as the Example-Based Machine Translation (EBMT). Neither syntactical nor semantic analysis is performed in such machine translation systems. Very large bilingual corpora are used instead and the translation process is performed on a higher level of word chunks. Word chunks in the source language are replaced with word chunks from the target languages. Appropriate word chunks substitutions are found in the bilingual corpora [5].

The Rule-Based Machine Translation systems represent a typical structural approach to the process of translation, while in the case of the Example-Based Machine Translation a cognitive approach to translation is taken. Both the above mentioned approaches have their merits and disadvantages. The author has developed his own new approach to machine translation which is called the Pattern-Based Machine Translation (PBMT). This approach is a combination of the rule-based and example-based approaches and tries to take the full advantage of the best features of these two approaches. In the Pattern-Based Machine Translation the translation process is performed not on the level of single words, but on the level of word chunks, However, contrary to the example-based approach, some kind of syntactical analysis in the pattern-based approach is still conducted [6].

The most important component of the Pattern-Based Machine Translation system is a database in which translation patterns are collected. These translation patterns play a crucial role in the whole translation process. There are three kinds of translation patterns: declination, conjugative and non-inflexion type patterns. The form of translation patterns is determined by the system of the target language. In the system developed by the author the target language is Polish which is a highly inflectional language with seven cases, three genders, and two grammatical numbers by which nouns, pronouns, verbs, and adjectives are declined or conjugated.

The declination translation patterns are the ones by means of which Noun Phrases (NP) are translated. These noun phrases play the role of the subject or object in the sentence. The pseudo-code for the declination translation pattern has the following form:

```
if <INPUT> == <SOURCE> then
{
          if <CASE> == 1 then <OUTPUT> := <TARGET_1>;
          if <CASE> == 2 then <OUTPUT> := <TARGET_2>;
          if <CASE> == 3 then <OUTPUT> := <TARGET_3>;
          if <CASE> == 4 then <OUTPUT> := <TARGET_4>;
          if <CASE> == 5 then <OUTPUT> := <TARGET_5>;
          if <CASE> == 6 then <OUTPUT> := <TARGET_6>;
          <CASE> := <CASE_TARGET>;
```

```
        <NUMBER> := <NUMBER_TARGET>;
        <GENDER> := <GENDER_TARGET>;
}
```

The field <SOURCE> is a phrase in the source language that is to be translated. Similarly, the fields <TARGET_1>, <TARGET_2>, <TARGET_3>, <TARGET_4>, <TARGET_5>, and <TARGET_6> are phrases in the target language, *i.e.* Polish, that are translations of the phrase in the field <SOURCE>. They are the declination forms for six cases of the Polish language (the vocative case is not taken into account, because it is an independent case and it is translated in a different was using non-inflexion type patterns). The fields <CASE_TARGET>, <NUMBER_TARGET>, and <GENDER_TARGET> are optional and they represent the values of such attributes as case, number and grammatical gender that are related to the Polish translation of the phrase in the field <SOURCE>. An example of the declination translation pattern can look like this:

```
if <INPUT> == ‘a solid base for’ then
{
        if <CASE> == 1 then <OUTPUT> := ‘mocna baza do’;
        if <CASE> == 2 then <OUTPUT> := ‘mocnej bazy do’;
        if <CASE> == 3 then <OUTPUT> := ‘mocnej bazie do’;
        if <CASE> == 4 then <OUTPUT> := ‘mocną bazę do’;
        if <CASE> == 5 then <OUTPUT> := ‘mocną bazą do’;
        if <CASE> == 6 then <OUTPUT> := ‘mocnej bazie do’;
        <CASE>:= 2;
}
```

The conjugative translation patterns are used mainly to translate Verb Phrases (VP). The framework for the conjugative translation patterns has a similar form to the declination translation pattern and is presented below in the pseudocode form:

```
if <INPUT> == <SOURCE> then
{
        if <NUMBER> == 1 and <GENDER> == 1 then
                <OUTPUT> := <TARGET_11>;
        if <NUMBER> == 1 and <GENDER> == 2 then
                <OUTPUT> := <TARGET_12>;
        if <NUMBER> == 1 and <GENDER> == 3 then
                <OUTPUT> := <TARGET_13>;
        if <NUMBER> == 2 and <GENDER> == 1 then
                <OUTPUT> := <TARGET_21>;
        if <NUMBER> == 2 and <GENDER> == 2 then
                <OUTPUT> := <TARGET_22>;
        if <NUMBER> == 2 and <GENDER> == 3 then
                <OUTPUT> := <TARGET_23>;
        <CASE> := <CASE_TARGET>;
        <NUMBER> := <NUMBER_TARGET>;
        <GENDER> := <GENDER_TARGET>;
}
```

The fields <TARGET_11>, <TARGET_12>, <TARGET_13>, <TARGET_21>, <TARGET_22>, and <TARGET_23> are the Polish translations of the phrase in the field <SOURCE> that are inflected according to the grammatical numbers and genders. An example of the conjugative translation pattern may have the following form:

```
if <INPUT> == 'failed to build' then
{
        if <NUMBER> == 1 and <GENDER> == 1 then
                <OUTPUT> := 'nie zdołał utworzyć';
        if <NUMBER> == 1 and <GENDER> == 2 then
                <OUTPUT> := 'nie zdołała utworzyć';
        if <NUMBER> == 1 and <GENDER> == 3 then
                <OUTPUT> := 'nie zdołało utworzyć';
        if <NUMBER> == 2 and <GENDER> == 1 then
                <OUTPUT> := 'nie zdołali utworzyć';
        if <NUMBER> == 2 and <GENDER> == 2 then
                <OUTPUT> := 'nie zdołały utworzyć';
        if <NUMBER> == 2 and <GENDER> == 3 then
                <OUTPUT> := 'nie zdołały utworzyć';
        <CASE> := 2;
}
```

The last translation pattern type is the non-inflexion type pattern which can be described by the following pseudo-code:

```
if <INPUT> == <SOURCE> then
{
        <OUTPUT> := <TARGET>;
        <CASE> := <CASE_TARGET>;
        <NUMBER> := <NUMBER_TARGET>;
        <GENDER> := <GENDER_TARGET>;
}
```

In this case the field <TARGET> represents the Polish translation of the source language phrase in the field <SOURCE>. An example of the non-inflexion type translation pattern is presented below:

```
if <INPUT> == 'up to this moment' then
{
        <OUTPUT> := 'jak dotychczas';
}
```

3. Fully-automatic machine translation system

The purpose of the specialized machine translation system which has been developed by the author is to translate precisely reports from the global foreign currency exchange market – forex. These reports are published twice a day on the Web site (http://www.mataf.net) and they concern 12 major currency pairs (mostly US Dollar and Euro crosses). The above mentioned reports describe the current situation from the perspective of the technical analysis for each currency pair. Moreover, information is given about the current trend, the configuration of some oscillators and the Bollinger bands. Moreover, some recommendations concerning trade opportunities for selected currency pairs are also given. It is recommended whether it would be profitable to open a short or long position on a given currency pair. In this case also some additional information is always given, *e.g.* where to put a stop loss and take profit orders. At the end of each recommendation assessment of the risk level of the recommended trade is also made, *i.e.* whether it is speculative to a certain degree or trend following.

On 4th March 2009 a report containing a forecast for the 12 major currency pairs was published on the Web site (http://www.mataf.net). For the purpose of presenting the results of the Pattern-Based Machine Translation system's work, a forecast for the EUR/USD currency pair was selected:

1,2555. EUR USD is in an downtrend directed by 1H exponential moving averages. EUR USD is in a consolidation after the last bearish movement. The volatility is high. Bollinger bands are flat. ForexTrend 1H, 4H, daily (Mataf Trend Indicator) is in a bearish configuration. The price should find a resistance below 1,2600 (45 pips). The downtrend should continue to gather momentum. We could take a short position at 1,2550. We will put the stop loss above 1,2620 (-70 pips). The targets are 1,2420 (+130 pips, risk/reward 1:1.9), 1,2350 (+200 pips, risk/reward 1:2.9). Each trade is dangerous, take care and put your stop loss. Trade configuration (1 Speculative → 4 Trend following): 3.

The specialized machine translation system translated the above quoted financial report into Polish in the following way:

1,2555. Para walutowa EUR/USD jest w trendzie spadkowym wyznaczonym przez jednogodzinne wykładnicze średnie kroczące. Para walutowa EUR/USD znajduje się w konsolidacji po ostatnim ruchu spadkowym. Zmienność jest wysoka. Wstęga Bollingera jest płaska. Wyznacznik trendu dla rynku forex obliczony dla danych godzinnych i czterogodzinnych, obliczony w horyzoncie czasowym jednej doby (indykator trendu biura maklerskiego Mataf) wskazuje na rynek niedźwiedzia. Cena powinna natrafić na opór poniżej poziomu 1,2600 (45 punktów). Trend spadkowy powinien przybrać na sile. Możemy zająć pozycję krótką na poziomie 1,2550. Ustawilibyśmy zlecenie obronne powyżej poziomu 1,2620 (-70 punktów). Poziomy docelowe są 1,2420 (+130 punktów, stosunek zysku do ryzyka 1:1.9), 1,2350 (+200 punktów, Stosunek zysku do ryzyka 1:2.9). Zajmowanie jakiejkolwiek pozycji jest niebezpieczne, należy zachować ostrożność i należy ustawić odpowiednie zlecenia *obronne. Ocena ryzyka zawieranej transakcji (od spekulatywnej: 1 do podążającej z trendem: 4): 3.*

4. Conclusions

A specialized machine translation system which was developed for the purpose of fully-automatic translation of reports coming from the global currency exchange market is presented in the paper. The obtained translation quality of financial reports is so high that the Polish version of these reports can be published without any further intervention of human translators. Moreover, the experiments conducted by the author with the help of English philology students from the Pedagogical University in Cracow have proved that a human translator who is not a specialist in the field of financial markets is not able to translate these financial reports better than the specialized machine translation computer program did. This fact shows that in the case of translation of great amounts of special kinds of documents it really pays to develope specialized machine translation systems that can perform the translation process faster and far better than human translators would be able to do.

The experimental machine translation system which has been presented in the article provides evidence that fully-automatic machine translation is possible. However, a lot of further research work should still be conducted in this field. Moreover, one must remember that research in the field of machine translation has sense only if it is oriented on a selected topic of texts to be translated [7]. Probably, it is not possible to built a machine translation system that could translate any kind of text with sufficient quality irrespective of the topic. The conclusion is that specialization of machine translation systems is absolutely necessary to achieve high quality of texts translated by computers.

Summarizing the paper it can be said that the key to success in implementing a fully-automatic machine translation system to forecast the behavior of financial markets is to restrict the subject of the translated text only to one specific thematic field. This fact will allowe all the necessary translation patterns to be collected in appropriate linguistic databases which will guarantee the sufficiently high quality of translated texts. Moreover, the Pattern-Based Machine Translation approach makes it possible to observe all the inflectional rules of the Polish language in such a way that the subject always agrees with the verb in terms of the grammatical gender and number. Moreover, the object is always used in the correct grammatical case.

References

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

[2] Hutchins W J 1986 *Machine Translation – Past, Present, Future*, Ellis Horwood Series in Computers and Their Applications, London

[3] Whitelock P and Kilby K 1995 *Linguistic and Computational Techniques in Machine Translation System Design*, UCL Press, London

[4] Gajer M 2002 *Machine Translation Review* **13** 7

[5] Gajer M 2004 *Control and Cybernetics* **34** (2) 357

[6] Gajer M 2008 *The Multilingual Pattern-Based Machine Translation Systems*, AGH University of Science and Technology Press, Cracow

[7] Melby A 1999 *Machine Translation Review* **9** 6