Machine Translation Errors: A Philosophical Examination

Mihail Radu Solcan
University of Bucharest
Department of Philosophy

Last Revision Date: 19 May 2006

 Version 1

Abstract

The paper treats machine translation as an experiment. It examines some types of negative results produced by this experiment. The paper starts with an analysis of machine translation systems. The components of these systems embody sophisticated natural language processing methods. They could cope in principle with obstacles raised by meaning and the knowledge of the world. The paper concludes that the main obstacle is the inability to read the mind of the author of a text. This should not be a problem if the human mind had a linguistic structure. The most plausible conclusion seems to be that the mind's structure is not linguistic.

1. Introduction

I will discuss first two open source machine translation projects. Then I will sketch three approaches to machine translation.[1]

1.1. The rarity of open source machine translation software

Open source software would be of great value for the present investigation. The access to the sources of the machine translation programs gives us an insight on what the pro-

[1] I ignore the statistical approach since I have a very limited knowledge of it. As far as I can guess the statistical approach is also less relevant for the basic argument of this paper.

gram can and cannot do. Otherwise we have to guess from experiments with the translation of fragments of texts.

Unfortunately open source machine translation projects are rare. I will present two such projects. They are from two distinct categories. The first is that of simple translation projects. The second is the category of sophisticated translation projects, backed by a research team.

- **The *Linguaphile* project**

Let us say that we have to translate the following sentence:

elle chante une belle chanson

from French to English.

I will describe what a system of translation in which words of one language are replaced by words of another

language and are adapted to the morphology of the target language. It is a simple system of translation. I am using Andrew Dunbar's *Linguaphile*[3], a command line machine translation program.

The first step is to identify grammatical words such as the article 'une'. One has also to identify the grammatical status of words such as 'chante'. I will call this phase "grammar and content separation".

The second step is to change the content words. For example, "chanter" and "chanson" were missing from the *Linguaphile* package downloaded from the Internet. I have added them and one can see quite easily how is built the lexical data-base:

```
'chanson'=> { 'x' => 'song',
              't' => 'n',
```

```
            'g' => 'f' },
 'chanter'=> { 'x' => 'sing',
            't' => 'v' },
```

I will call the second phase "content transformation".

In the third phase the computer program has to do things like the transformation of 'sing' into 'sings'. This is done because, in French, 'chante' is the form of the present third person singular of the verb. I will call this phase content and grammar unification (in the target language).

Now, the commands for translation in *Linguaphile* are quite straightforward:

```
./linguaphile.pl -s fr
** translating from french to english:
```

```
elle chante une belle chanson
she sings a beautiful song
```

It is obvious that in the case of the above translation grammatical structures are quite similar. See what happens however when we vary a bit the syntax of Romanian, making it less similar to that of English:

```
./linguaphile.pl -s ro
** translating from romanian to english:

ce trimite Paul la universitate?
what send Paul at university?

Paul trimite scrisoarea la universitate
Paul send the letter at university
```

The lack of the syntactic structures makes room immediately for errors in translation.

- **The Apertium project**

The *Apertium* is much more sophisticated than *Linguaphile*. The project has been developed at the University of Alicante.[2]

The *Apertium* translator uses a translation method called by the authors "shallow transfer". This method is suited for closely related languages, such as Spanish, Catalan or Portuguese.

[2]The sources of Apertium[1] are available on the Internet. They include a description of the system and an user's manual.

Let us see a series of examples. First an example with a text in Portuguese.[3]

> Em linhas gerais, Platão desenvolveu a noção de que o homem está em contato permanente com dois tipos de realidade: os inteligíveis e os sensíveis.

On the command line we ask for a translation:

```
apertium-translator apertium/es-pt pt-es txt
< platon.txt > platon-es.txt
```

The result is:

> En líneas generales, *Platão desarrolló la noción de que el hombre está en contacto perma-

[3]The text is taken from the article on Plato in the Portuguese version of Wikipedia.

> nente con dos tipos de realidad: los *inteligíveis y los sensibles.

The unknown words are marked with an asterisk. It is even possible to reverse the translation:

> Em linhas gerais, **Platão desenvolveu a noção de que o homem está em contacto permanente com dois tipos de realidade: os **inteligíveis e os sensíveis.

As I have a rather superficial knowledge of Spanish and Portuguese it is difficult for me to assess the accuracy of the results in more intricate situations. However, the translator copes quite well with such sentences:[4]

[4]The sentence is taken from the article on Plato in the Spanish Wikipedia.

> A Sócrates lo menciona frecuentemente en los diálogos.

This becomes in Portuguese:

> A Sócrates menciona-o frequentemente nos diálogos.

It is even possible to translate the following conclusion of an article about Searle's Chinese room:[5]

> De acuerdo a los creadores del experimento, los defensores de la inteligencia artificial deben admitir que, o bien la sala comprende el idioma chino, o bien el pasar el test de Turing no es prueba suficiente de inteligencia.

The text is translated into Portuguese as:

[5] In the Spanish Wikipedia at http://es.wikipedia.org/wiki/Sala_china.

> De acordo aos criadores do experimento, os defensores da inteligência artificial devem admitir que, ou bem a sala compreende o idioma chinês, ou bem o passar o teste de *Turing não é prova suficiente de inteligência.

The advantage of the open source software is that we are really able to look into the "Chinese room" (in this case, the "Spanish-Portuguese room"). According to the documentation included in the Apertium software package[1], the system is made up of a chain of modules. I will ignore the modules which deal with the html tags or the rtf format commands. They are just auxiliary modules. They play no role in the case of text files.

The first modules are a morphological analyzer and a tagger. Then, in the core of the chain of modules, there

are the modules for structural and lexical transfer. At the end there are the modules for morphological generation and post-generation of the text in the target language.

Each module works independently. For example the sentence about Plato mentioning Socrates is analyzed morphologically with the help of the following command:

```
lt-proc -a es-pt.automorf.bin
< socrates.txt > socrates-morph.txt
```

The result is the following:

```
^A/A<pr>$
^Sócrates/Sócrates<np><ant>$
^lo/lo<detnt>/lo<prn><pro><p3><nt>/lo<prn>
     <pro><p3><m><sg>$
^menciona/mencionar<vblex><pri><p3><sg>/mencionar
```

```
     <vblex><imp><p2><sg>$
^frecuentemente/frecuentemente<adv>$
^en/en<pr>$
^los/el<det><def><m><pl>/lo<prn><pro><p3><m><pl>$
^diálogos/diálogo<n><m><pl>$
^./.<sent>$
```

The machine translation is the result of a chain of actions performed by modules. Each module changes the text until it is reached the target language.

- **Machine Translation as Natural Language Processing**

Machine translation, as an endeavor to use computers for translating from one natural language to another natural language, is part of the area of natural language process-

ing.[6]Computational linguistics and even speech recognition may be involved in this process.

An important point for the final argument in this paper is the relevance of the use of modules in machine translation as discussed in section •. Each module does a certain operation. Each operation represents some form of processing a text in a natural language.

There are various toolboxes for natural language processing. I will focus again on open source software.

Steven Bird, Edward Loper and others have developed the Natural Language Toolkit(NLTK).[7]This is a suite of programs written in a computer language called Python.

---

[6]See Chapter 21 in Jurafsky and Martin's book[5].

[7]NLTK has a web page with a lot at http://nltk.sourceforge.net/.

Another toolkit is FreeLing. This is also a suite for the analysis of texts written in a natural language. FreeLing is written in C++.[8]

As an illustration I will use an analysis with the help of FreeLing. I use the following text (from the documentation of Apertium):

El gato come pescado.

The command

```
analyzer -f es.cfg <gato.txt >gato.mrf
```

generates the following file:

```
El el DAOMS0
gato gato NCMS000
```

[8]FreeLing has a site at http://garraf.epsevg.upc.es/freeling/.

```
come comer VMIP3S0
pescado pescado NCMS000
. . Fp
```

It is easy to see that the analyzer has detected that *come* is a form of the verb *comer*.

The flux of the machine translation process is made up of operations that resemble the above analysis. Of course, the text might be much more ample. But the idea is to keep each operation simple and to put them together in a final chain, having at the beginning the source text and at the end the target text.

1.2. Three approaces to translation

The approaches sketched here are inspired by the presentation in Jurafsky and Martin's book[5]. I have however changed the order of presentation.

- **Direct translation**

The three phases approach described in section • is somehow simplistic.[9]

However, Hutchins and Somers characterize direct translation as a chain of three modules.[10]

Research from the 1950's, such as Kenneth Harper's study

[9] Nigel Ward, Jurafsky and Martin[5, p.816] offer a six phases description of direct translation from Japanese to English.

[10] Hutchins and Somers[4, p.132].

of Russian, exhibit the tendency to adopt this form of approach to translation.[11]

The advantage of direct translation is that it always yields some result.

The disadvantage of direct translation is that it has almost no chance to work in the case of ample differences in the syntax of the source and target languages.

- **Transfer**

The basic idea of transfer is to include in the translation chain a syntactic parser. This creates however a problem.

If the syntaxes of the two languages are very different there is a gap between the source language parse tree and

[11] See Harper's paper in Locke and Booth[6, pp.66–85].

the target language parse tree. The solution is to fill the gap with a syntactic transformation of the first tree into the second tree.[12]

- **Interlingua**

Sometimes, such requirements as the translation into all the official languages of the European Union, lead to a different approach. A form of meaning representation, called *interlingua* is used.[13]

The three approaches to machine translation are represented graphically with the help of a pyramid diagram.[14]Direct

[12]For details see Nigel Ward's explanations[5, p.808].
[13]For details see Ward's presentation[5, p.812].
[14]See the diagram in Hutchins and Somers[4, p.107].

translation forms the base of the pyramid. On top is interlingua.

2. Understanding meaning

Most of the available translation software is commercial software. I did some experiments with commercial software. I will mention the source only in the case of replicable experiments. These are experiments with online available software.

The errors are sorted from a human perspective. Human translators also make mistakes. They might, for example, misunderstand the meaning of a word. The result is a translation error that one may often encounter.

2.1. Title translation

The translation of titles is sometimes a difficult task even for experienced human translators. I will start however with a very simple title. No human translator would probably make a mistake in this case, unless she or he simply looks up the words in the dictionary and makes no effort to "catch the meaning".

> Che cosa ha veramente detto Gramsci?[15]
>
> Machine translation:
>
> What does it have really motto Gramsci?
>
> Human translation: What did Gramsci really say?[16]

[15] Mentioned in Allen[2, p.231].

[16] Allen[2, p.231].

The word ‘detto’ has at least three meanings in Italian. First, ‘detto’ is the past participle of ‘dire’(say). Second, it might be used phrases such as ‘Leonardo Bruni, detto l’Aretino’. Third, it has more or less the meaning of *motto*.

The problem of homophones which have different meanings or grammatical uses is frequent in many languages. I have no idea how the commercial software used in the experiment tries to cope with homophones. Apertium[1] uses a module called ‘tagger’ in order to deal with homophones.

I did some experiments with variations of the above Italian text. First, I eliminated the name of Gramsci. Then I have changed the word order, as one can see below:

Che cosa ha veramente detto?

Machine translation:

What does it have really motto?

Che cosa ha detto veramente?

Machine translation:

What has you/he/she said really?

The computer program got the meaning of detto, but was caught in the ambiguity of the Italian sentence.

The grammar of the following Spanish text, according to Allen[2, p.169], is rather difficult. A human translator would try to catch the imaginative note in the following title:

El mundo es ansí[17]

Machine translation:

The world is ansí

[17] Apud Allen[2, pp.169-170].

Human translation: The world is like that[18]

This title created difficulties for Apertium[1] too. It reacted in the same way as the commercial software. The word ‘ansí’ is the popular form of the litterary ‘así’. With the litterary word the machine translation becomes less problematic, but I have changed the connotations of the original title.

2.2. Small vs short

I find the following experiment very interesting. Again, human translators would rely on their understanding of the meaning of the text. The computer program has however

[18] Allen[2, p.170].

problems when it has to choose the right meaning in the target language.

I use the title of a section in a german book:[19]

> Kleine Geschichte der Erkenntnistheorie
>
> Google translation:[20]
>
> Small history of the theory of knowledge

Altavista Babel Fish[21] returned the same result.

Now, the problem is that this is a "short history of the theory of knowledge". 'Small' is not appropriate in this context. Human translators might make however the same

---

[19] Herbert Schnädelbach, *Erkenntnistheorie zur Einführung* (Hamburg: Junius, 2002).

[20] http://www.google.com/language_tools?hl=en

[21] http://babelfish.altavista.com/

kind of mistake. For example, in Romanian, a language for which I have a better feeling for the 'small-short' distinction, one might say "mică istorie a teoriei cunoaşterii", instead of "scurtă istorie a teoriei cunoaşterii".

The choice of the right meaning is even for human translators a matter of fine distinctions.

2.3. Could the machine cope with meanings?

I think that the solution of the problems connected with meaning is not impossible. It all depends on a less syntax-centered translation and a more meaning-focused translation.

My intuition is that machine translation programs would return better results if it would be possible to construct

a data-base of the meanings in the source text. Perhaps, unlike in the interlingua approach, one should transform this data-base according to the habits of the target language and then generate the target text.

3. Interest in the world

Let us start again with the example of a title.

Cristo si è fermato a Eboli[22]

MT version:

Christ him and stopped to Eboli

Instead of

Christ stopped at Eboli[23]

[22] Mentioned in Allen[2, p.231].

[23] Allen[2, p.231].

The machine does not seem to display any sort of interest in Christ or for the book of Carlo Levi. Does this mean that the good news are not heard beyond the city of Eboli? It is difficult also for humans to grasp the meaning of the title without reading the book.

The translation software also seems to have problems with homophones. A human translator would however solve the problem "going beyond language" and trying to find out who is Christ or why would he stop at Eboli.

In the following example one has to know something about the arms of Ottoman soldiers:

> Wie der Türke auf der Karlsbrücke um seinen Säbel kam[24]

[24] Apud Allen[2, p.30].

MT version:

How the Turk lost his/its saber on the Karl-bridge

Human translation:

How the Turk on the Charles bridge lost his scimitar[25]

In the next example it would be easier to know something about Montesquieu, rather than about laws:

Observations sur un livre intitulé: De l'esprit des lois

Machine translation:

Observations on a titled book: Of the law mind

[25] Allen[2, p.30]. The 'scimitar, scimiter' is a kind of Ottoman sword, called in Romanian 'iatagan'. See also http://en.wikipedia.org/wiki/Scimitar.

In fact, the rule would be to leave the title in French, if no translation is available in English. Otherwise, the title may be translated. But in order to do this the program has to know something about the world.

3.1. Kant's definition of Enlightenment in the vision of machines

I did some experiments with the machine translation of fragments from Kant's works. It is easy to guess that they are full of various mistakes. But the following text is not from Kant and it is not very difficult. However, there are many problems with the machine translations:

> Kant defined the Enlightenment as an age shaped by the motto, "Dare to know" (latin: Sapere

aude). This involved thinking autonomously, free of the dictates of external authority.[26]

Google translation into French:
Kant a défini l'éclaircissement comme âge formé par la devise, « défi pour savoir » (latin : Aude de Sapere). Ceci a impliqué de penser de façon autonome, exempt des préceptes de l'autorité externe.

Google translation into German:
Kant definierte die Aufklärung als Alter, das durch das Motto, „trauen geformt wurde, um zu wissen" (Latein: Sapere aude). Dieses bezog

[26]Modified version of a fragment from the article on Kant in Wikipedia.

mit ein, autonom zu denken, freigeben von den Vorschriften der externen Berechtigung.

The disaster in the german translation is emphasized by the following contrast with the actual German version of the same article on Kant: Sapere aude - Habe Mut zu wissen.

3.2. Possible MT solutions

It is difficult to say how one could emulate 'interest in the world' in a computer program. But, from the point of view of the knowledge of the world, the situation is not hopeless for MT.

The idea is to emulate the use of knowledge by human translators. In principle, it is possible to create data-bases

that would store much more information about the arms of Ottoman soldiers, Hitler and so on than a human mind.

Fast search in data-bases is now possible and this knowledge could be used in machine translation. It would certainly cope with the problem of the Ottoman scimitar.

4. Translation and the mind of the author

Human translators try to figure out what is "in the mind" of the author.[27] The following example is very simple if one thinks about a person who studies the language of Bogotá.

[27] The phrase "mind reading" is fairly usual in cognitive science. It seems to betray a tacit analogy between a mind and a text. The nature of this process is vividly discussed. See Robert M. Gordon, "Simulation vs. Theory-Theory" in Robert A. Wilson and Frank C.

This person might write some notes on her researches and so on.

> Apuntaciones críticas sobre el lenguaje bogotano[28]
>
> Commercial MT version:
>
> Apuntaciones critics on the language from Bogotá
>
> Human translation:
>
> Critical notes on the language of Bogotá[29]

The open source Apertium[1] copes better with the above example. It also lacks in the dictionary some words, but seems to take advantage of the close relationship with Por-

---

Keil(eds.), *The MIT Encyclopedia of the Cognitive Sciences* (Cambridge, Massachusetts: The MIT Press, 1999), pp.765–766.

[28] Apud Allen[2, p.169].

[29] Allen[2, p.169].

tuguese and produces the following translation into this language:

*Apuntaciones críticas sobre a linguagem *bogotano

In contrast to the first example, the following one is difficult to translate in languages that do not offer the possibility to replicate the pun. In German, 'Führer' means *leader* and 'Verführer' means *seducer*. The example reproduces the title of a German documentary film:

Hitler - Verführer des Volkes

Machine translation by a commercial software:

Hitler - seducers of the people

Google translation:

Hitler - leaders of the people

InterTran translation:[30] Hitler tempter of the people

PROMT translation:[31] Hitler - seducer of the people

The last two translations are correct, but they lack an insight into the mind of the author of the title. If we try to "read" the mind of the author we see Hitler as tempter and leader (into adventure and disaster).

As far as I know no machine translation software tries to read the mind of the authors of the text. In contrast, I think that a human translator understands in a fundamental way the mind of the author.

[30] http://www.tranexp.com:2000/Translate/result.shtml

[31] http://www.e-promt.com/

5. Conclusion

The first step of the closing argument in this paper is connected with the investigation of various aspects of natural language processing. As has been shown long ago by Terry Winograd[7] the computer approach to the investigation of the natural languages has brought in more, not less, rigor. The methods of computational linguistics are far more sophisticated than the traditional methods in linguistics.

Let us also adopt an optimistic view concerning the possibility to build semantic data-bases and to represent knowledge.

Why would not all these advances bring us closer to a model of the mind and then to the possibility to build programs which read the minds of the authors? According to

very popular views in cognitive science the mind has some sort of linguistic structure. Understanding language should bring us closer to the understanding of the mind.

Machine translation is however like a gigantic experiment which brings forward interesting negative results. There are areas in which machine translation unrelentlessly generates errors.

According to the argument developed here this raises doubts concerning the linguistic structure of the mind. Language is important for the human mind, but it does shape it. This would be the reason for the lack of the ability to read a mind if we treat it as a text or some other kind of object which has a linguistic structure.

References

[1] Carme Armentano-Oller, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Boyan Bonev, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, and Felipe Sánchez-Martínez. *Apertium: an open-source shallow-transfer machine translation engine and toolbox.* <http://apertium.sourceforge.net>, visited at 2006-05-11. Quoted at pp. 8, 12, 23, 25, and 35.

[2] C.G.Allen. *A Manual of European Languages for Librarians.* London: Bowker, 1975. Quoted at pp. 22, 24, 25, 28, 29, 30, and 35.

[3] Andrew Dunbar. *Linguaphile* [The web page of an easy to use command line translator]. <http://

`linguaphile.sourceforge.net`>, visited at 2006-05-11. Quoted at p. 5.

[4] W.John Hutchins and Harold L. Somers. *An introduction to machine translation*. London: Academic Press, 1992. Quoted at pp. 18 and 20.

[5] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Upper Saddle River, New Jersey: Prentice Hall, 2000. [Chapter 21, on machine translation, is mostly written by Nigel Ward]. Quoted at pp. 15, 18, and 20.

[6] William N. Locke and A. Donald Booth, editors. *Machine Translation of Languages*. Cambridge, Massachusetts: The MIT Press, 1955. Quoted at p. 19.

[7] Terry Winograd. *Language as a Cognitive Process*. Reading, Massachusetts: Addison-Wesley, 1983. Quoted at p. 38.