[From: *Tralogy*, Paris, 3-4 March 2011]

Sergey Kulikov

What is web-based machine translation up to?

Abstract

This paper deals with machine translation on the Web. In the first section we outline the problems of machine translation, paying special attention to the web-oriented systems. In the second section we present a method of using multidomain corpora for the proper style output. The third section turns to text-level analysis and briefly discusses anaphora and coreference resolution. The fourth section turns to other NLP systems that may improve the quality of machine translation on the Web. In conclusion, we set the range of goals for the nearest future.

Introduction

Since the very beginning of machine translation in 1954, it has focused on translation of scientific or technical documents. It was basically used in large translation agencies to translate large scale documents of a certain type, like nuclear reactors descriptions or aircraft manuals. A fast low quality translation with necessary post-editing was then considered the main purpose of machine translation. The systems were based on large dictionaries containing all the terminology available for a certain domain. After some years of using a machine translation system, it could be improved to give a satisfactory translation, provided that it has been used on a certain narrow domain. However, it was generally accepted that it took nearly the same time to post-edit a low quality machine translation output or to translate it manually. But with the introduction of the Internet in early 1990s machine translation became a necessity for millions of people. People needed a fast translation of web-pages, e-mails, and for the majority of them a 100% accurate translation was not a must. Therefore, machine translation could be used by them with no restrictions and it started to develop rapidly. Over the past 20 years the quality of machine translation has risen but some people have started to blame the output quality. However, Internet contains large amounts of multilingual data that is constantly changing, so no matter how large the team of translators may be it will never keep up with the speed of growth, change and update of the data in proportions that can be thought of as sufficient. In such a case we simply have to acknowledge that humans cannot rival machine translation on the Web.

Some problems of machine translation on the Web

Types of machine translation systems on the Web

Currently all the systems of machine translation can be divided into 3 types (depending on the approach they are based): 1) traditional rule-based systems, 2) statistical systems and 3) hybrid systems which combine to some extent the features of the first two types. We will focus on the first two types of machine translation systems as hybrid systems are either in the process of development or else do not play a significant part in the world of machine translation. However, there is another classification possible. According to it, machine translation systems can also be divided into 3 types: 1) systems that are integrated into web-browsers, 2) systems that are part of a multilingual search engine and 3) on-line translation services. There are no direct correspondences between these two classifications. But the systems integrated into web-browsers are mostly rule-based (with a notable exception for *Google Translate*©), whereas the systems that are part of a search engine are basically statistical (with an exception at least for the leader of Russian web-search market Yandex©). And on-line services are provided by both approaches.

There are some similarities between rule-based and statistical machine translation systems if we consider their operation on the Web. When speaking about rule-based MT we usually bear in mind that we can range the system's dictionaries according to our understanding of the topic of the text. And we need to have a training corpus of a certain domain for a statistical MT system. But while using any type of MT on the Web we can neither specify the set of dictionaries nor the training corpus.

Types of Web sites as a source of MT mistakes

The Internet has several different types of information representation. Among the most common types are: *news-line type*, *news type*, *list type*, *hybrid type*, *blog type*, *plaintext type*, *website (structure)* and *e-mail type*. All these types can also be subdivided according to the implied audience.

News-line type is a type of news representation where news is presented by their titles only. This type is common on news websites (like BBC website, or Yandex.News).

Московскиешколызакроютнакарантиниз-загриппа (Russian. Yandex.News ©)
The Moscow schools will close on quarantine because of a flu (Promt ©)
Moscow schools will be closed to quarantine because of the flu (Google ©)
Moscow schools to be closed in quarantine: Flu to blame. (Normal News Headline)

It can be seen that neither rule-based Promt nor statistical Google make use of the interpretation type. Therefore, while the output is more or less correct, the style is improper. The style is very similar to newspaper headlines. The sentences tend to be short and have simple grammar. Of the two systems Google provides a better translation.

News type is similar to the previous one. It consists of a headline and a short synopsis. Thistypeisalsocommonnewswebsites.

На курортах Египта остаются около 20 тысяч россиянПо данным внешнеполитического ведомства, с конца января из Египта выехали более 20 тысяч

россиян, сейчас на курортах Хургады и Шарм-аш-Шейха остаются около 28 тысяч туристов из России. (Russian. Yandex.News ©)On resorts of Egypt remain about 20 thousand Russians
According to foreign policy department, from the end of January Egypt have left more than 20 thousand Russians, now on resorts of Hurgada and the Charm-ash-sheikh remain about 28 thousand tourists from Russia. (Promt 9.0 ©)
The resorts of Egypt are about 20 thousand Russians
According to Foreign Ministry, in late January from Egypt left more than 20,000 Russians, now in the resorts of Hurghada and Sharm el-Sheikh are approximately 28 thousand tourists from Russia. (Google ©)
About 20 thousand Russians still on Egypt resorts
According to Ministry of Foreign Affairs data, more than 20 thousand Russians have left Egypt since late January, now nearly 28 thousand tourists from Russia are still left at the resorts of Hurghada and Sharm el-Sheikh. (Human Translation)

If we compare the translations the main problems would be word order and name entity translations.

List type is a way of information presentation common on search engines. In some aspects it is nearly the same as the previous type. The main difference is that instead of a synopsis this type has a set of the inquiry mentions in the document. The problems are the same – word order and name translation.

Plain text type is a way of presenting long text documents in one of the web formats (such as html, xml, txt). We will discuss this type in the third section.

Blog type sites are among the most visited ones on the Web, so their high-quality translation is a must for the near future. They consist of a headline, text and comments.

Alles alternativlos!
Wenn´s brennt, muss eben gelöscht werden. Alternativlos. Und weil es in unserem System überall brennt, sind unsere Handlungsentscheidungen in Folge auch alternativlos. Punkt. Wenn Staaten oder Banken sich nur so hoch verzocken, dass ihr Versagen ganze Volkswirtschaften mit in den Strudel zieht, dann müssen wir sie eben auch tot noch am Leben erhalten. Alternative? Fehlanzeige. (German, http://www.werbeblogger.de/)
All alternative!
If it burns, must be deleted indicated. Alternative. And because it burns anywhere in our system, our choices of action in a row, even without alternative. Point. If governments or banks gamble only so high that their failure all economies is taken into the whirlpool, then we must also just dead nor alive. Alternative? Not applicable. (Google ©)
All without alternatives!
If´s burns, just must be deleted. Without alternatives. And because it burns in our system everywhere, our action decisions in consequence are also without alternatives. Point. If countries or banks stretches only so highly verzocken, that its failure entire economic systems with in the swirl, we must receive it just also dead yet alive. Alternative? Wrong. (SDL Trados ©, http://www.freetranslation.com)

Everything alternativlos!
if it burns, must be just extinguished. Alternativlos. And because it burns in our system everywhere, are our action decisions as a result of also alternativlos. Sharp. If states or banks to ourselves only so high verzocken that her failure pulls whole national economies with in the whirlpool then we must still keep alive them just also deadly. Alternative? Dead loss. (Promt ©, www.translate.ru)
There`s no alternative!
If something is on fire, it must be put off at once. No alternatives. And as our system is alltogether on fire, we therefore have no other choices left. Point. If governments or banks gamble so much that their failures land the whole economy in a whirlpool, then we must consider them just as dead yet alive. Alternative? No chance. (Human translation)

We can see that this information representation type uses colloquial patterns and meanings, thus, it is similar to some extent to speech or very informal style. Here MT can provide us only with a very vague idea of actual meaning of the author.

E-mail type can vary due to the type of the letter. Formal letters differ greatly from informal ones in form, style, use of grammar, etc.

Web sites usually have a certain set structure. Their main page can be named in several different ways – homepage, home, main, etc. In general, these titles are translated properly by MT systems.

Hybrid type combines some of the features of the previous types. Moreover, photohostings having text descriptions are also considered this type.

We can see that type of the document influences the quality of MT due to both pragmatic purposes and style of presentation.

The same results are also true for scientific and technical translation where a stylistic typology of texts for machine translation purposes is being developed [Marchuk, Yu. (2010)].

Corpus linguistics and automatic domain-detection

Characteristics of modern corpora

Nowadays corpus linguistics has made its way not only into natural language processing but into traditional linguistics as well. Corpora can be national consisting of at least 100 million words, specialized i.e. representing a certain domain, parallel i.e. having aligned texts from two different languages, etc. Most corpora are annotated which makes them useful for various kinds of research.

Types of annotations

The most frequent annotation is morphological or part-of speech annotation. It does not only provide valuable information about the text but also serves as an initial step for further analysis. Another common type of an annotated corpus is a syntactically anotated corpus (also called parsed corpus or treebank).

But if the principles of the basic annotation types are widely accepted and are of more or less the same standard, the less common annotation types are often presented in very different formats. These annotations include anaphora, discourse, pragmatic, sentiment and some other annotation types.

Parallel corpora are most widely used for the purposes of both human translators' training and statistical machine translation. While they are not generally annotated, they can be aligned in two different modes – sentence alignment and word alignment.

Basic parts of a linguistic corpus

A large (national) corpus generally consists of several parts – a corpus manager, texts, annotation and subcorpora, i.e. collections of texts that are connected by some common feature – the same author, the style, the year of publication, etc. It should be noted that the division of a corpus into subcorpora can also be achieved by the corpus user.

Most of national corpora have an Internet subcorpus in them. However, it is generally insignificant considering the size of a corpus.

Parallel corpora

Parallel corpora are as has been mentioned before of two types – sentence aligned and word aligned. The is a third type, which is the base of phrase-based statistical MT, – n-grams which are often refered to as phrases, though from a purely linguistic point of view these n-grams can be anything, like *out of the*. Word alignment is achieved automatically. The results, depending on the type of the document and the language pair, can vary from relatively poor in case of sentence aligned fiction corpus to nearly 100% accurate in case of a word aligned bilingua text or a highly formalized official document.

The basic problems with parallel corpora are limited number of such corpora, at least for some language pairs, insignificant number of different domains, especially related to the Web and lack of linguistic annotation. Although several methods to cope with some of these problems have been suggested [Koehn, P. (2009)], such as using comparable corpora that can be acquired directly from the Web, creating one's own parallel corpus using data from multilanguage sites, etc.

Parallel corpora and automatic domain-detection

As we have shown in the previous section, automatic domain detection could improve the quality of Web-based MT significantly. But creating corpora in which one domain will be clearly distinguished from the other is extremely difficult. As it has been shown by long

years of studying both terminology and machine translation output this task cannot be achieved by using lexical information only, since terminology seems to overlap in different domains, sometimes with a shift in meaning. Yet since web-based MT has to deal with all the domains present in the Internet we cannot determine so easily terminology from how many domains is present in a document, mainly due to the fact that either there is no terminology at all, or it will not help us that much.

We propose that each of the information representation types if we consider them as a separate domain should be syntactically analyzed so as to create a special treebank. At the first stage such corpora should be made for non-parallel texts for as many languages as possible in order to find out surface syntactical structures that are considered natural by the native speakers for this domain. Thus we will be able to determine the stylistical and syntactical features that are typical for each domain. Making treebanks is not a very difficult task any more. Since "the Statistical Revolution" of the late 1980s there have appeared a number of different parsers, both free and commercial, which are available for all major languages while some of them are language-independent. After that parallel corpora of these domains can be created.

The process of automatic domain-determination in a MT system that we propose is the syntactical analysis of the input text. A further improvement in automatic domain identification may be achieved by combining url-analysis with the proposed method. In case of the system's output we have to use the information from the above mentioned parallel corpora.

We can use the same technique for MT in general. If we use syntactically annotated parallel corpora, then we can preserve the communicative structure of the output text. This will, however, make the language model used by statistical MT systems and make the systems work a little slower but we believe that the increase in translation quality will compensate for the time.

How to achieve coherent text output

Incoherent text output is an old problem of MT. While most systems do well at the sentence level, they fail to achieve the same good result at a paragraph level, to say nothing of the whole text.

Traditionally, in case of machine translation, the problem of coherent output meant correct anaphora resolution and proper dictionary selection to get the right terminology, but with the introduction of statistical MT the situation has greatly changed. An anaphora resolution component is where lies yet another difference between rule-based and statistical machine translation. Rule-based systems tend to have such a component while there is no such a thing in statistical MT.

The reason why statistical MT systems do not have an anaphora resolution component comes from the very structure of the systems. In brief, they consist of a training corpus, a training (statistical machine learning) model and a target-language model [Koehn, P.

(2009)]. So there is no information that is necessary for anaphora resolution present in the system, since this information relies on the input text.

What makes a text coherent?

Anaphora had been the only discourse relation that was analyzed in NLP until mid1990s[1]. Even though there were influential works on ellipsis and lexical cohesion. The problem was that an anaphoric relation can be easily found and without a proper antecedent it is nearly impossible to get a right translation of pronouns. This is especially important for language pairs with different typological structure. The English pronoun "they" has no gender, however when translating into, for example, Spanish, we need to know the gender of the antecedent, as the Spanish equivalent is different for the masculine (*ellos*) and feminine (*ellas*). If we take the English pronoun "both", which also lack gender, then in order to get a right Russian translation we need to find out whether "both" refers to two men, two female or a man and a woman.

Since mid1990s another broader view upon anaphora has prevailed. The study of anaphora in a narrow sense (i.e. third person singular anaphora) was substituted by studies in coreference resolution. These studies were carried out on text corpora which enabled to create better systems for both anaphora and coreference resolution.

But anaphoric and coreferential relations are not the only type of cohesion. There are also lexical and grammatical synonymy, discourse particles, stylistics and other less known means of cohesion.

Machine translation of a coherent text

It seems likely that machine translation since it has only an anaphora resolution component (in case of a rule-based system) does not make use of all the information about means of cohesion in text. In order to prove the above statement we present a short study of cohesion below. By numbers we indicate various text relations referring to the some part of the text. We mark the ambiguous cases where more than one potential reference can be chosen by brackets {}.

Aragorn 1, Legolas 2 und Gimli 3 treffen zu dieser Zeit Gandalf 4 wieder 5, der 4 ihnen 1, 2, 3 erzählt, dass er 4 zwar im Kampf 6, 4 gegen den Balrog 7, 6 gestorben 8, 4, 6 sei, nachdem 8 sie 9, 4, 7 sich 9, 4, 6 von den tiefsten Bereichen 10, die 10 weit unter den tiefsten Stollen 11 Morias 12 liegen, über die sagenumwobene 3 endlose Treppe 13, die 14 von den tiefsten Bereichen 10, 12 Morias 12 bis zu den Gipfeln des Nebelgebirges 14, 12 führt 13, bis zum Gipfel der Berge 15, 11 hochgekämpft 6, 4, 7 hatten 16, 8, wo 15 Gandalf 4 den Balrog 7 zwar erschlug 6, 8, 7, aber den Kampf 6 selbst 4 nicht überlebt 4, 8 hatte 16. Allerdings 4, 5 wurde er 4 „zurückgeschickt" 4, 8, da seine 4 Aufgabe 17, 4, Sauron zu stürzen 17, noch nicht erfüllt 17 sei. Er 4 beruhigt 18 die drei Gefährten 1, 2, 3 über das Verbleiben der Hobbits 19, in dem 18 er 4 ihnen 1, 2, 3 kurz erzählt 18, 19, wie 19 es ihnen 19 in Fangorn erging 19, und reitet dann mit ihnen 1, 2, 3 nach Edoras, damit Aragorn 1, Legolas 2 und Gimli 3 ihr 1, 2, 3 Versprechen 1, 2, 3 gegenüber Éomer

einlösen können.[2] Aragorn 1, Legolas 2 and Gimli 3 meet at this time Gandalf 4 again which 4 tells them 1, 2, 3 that he 4 has died 5 though in the fight 6, 4 against the Balrog 7, 6, after she 8 conducts herself 8 from the deepest areas 9 which 9 far lie under the deepest tunnel Morias 10 about the sagenumwobene endless stair, from the deepest areas Morias 10 up to the summits of the Nebelgebirges 11, up to the summit of the mountains 12, 11 highly fought 6, 4, 7 had where 12 Gandalf 4 killed the Balrog 7 though, but had not survived 5 the fight 6 even. Indeed, he 4 was "sent back" 4, 5, because his 4 job 13, 4 to overthrow Sauron 13 would not be fulfilled yet. He 4 calms three companions 1, 2, 3 about remaining the Hobbits in which he 4 briefly tells them 1, 2, 3 how it went out to them 1, 2, 3 in Fangorn, and then rides with them 1, 2, 3 to Edoras, so that Aragorn 1, Legolas 2 and Gimli 3 can redeem her promise towards Éomer. (Promt Online ©, www.translate.ru)
Aragorn 1, Legolas 2 and Gimli 3 encounter to this time Gandalf 4 again, that 4 tells them 1, 2, 3, that it 5 had died 6 to be sure 5 in the battle 7 against the Balrog 8, after it {5, 7} of the deepest areas 9, that 9 far under the deepest Morias 10 stood lie, high fought had over the legend enveloped endless stairway 11, that 11 leads of the deepest areas Morias 10 to the summit of the fog mountains 12, to the summit of the mountains 13, 12 where 13 Gandalf 4 killed the Balrog 8 to be sure, but that Battle 7 even survived 4 had not. To be sure it 14 became not "sent back" because its 14 task 15 to fall, Sauron, yet fulfilled 15 would be. It 14 calms the three companions 1, 2, 3 over the remaining of the Hobbits, in which it 14 tells them 1, 2, 3 shortly, how it 14 issued them 1, 2, 3 in Fangorn, and rides then with them 1, 2, 3 after Edoras 15 so that Aragorn 1, Legolas 2 and Gimli 3 can redeem its 15 promise vis-à-vis Éomer. (SDL Trados ©, http://www.freetranslation.com)
Aragorn 1, Legolas 2 and Gimli 3 meet Gandalf 4 again at that time, who 4 tells them 1, 2, 3 that he 4 had indeed died 5 in the fight 6, 4 against the Balrog 7, 6 after it 1 from the deepest areas 8 that 8 are far below the deepest tunnels of Moria
9, the legendary endless stairs the deepest areas of Moria 9 to the peaks of the Misty Mountains 10, out, had fought to the high peaks of the mountains 11, 10, where 11 Gandalf 4 slew the Balrog 7, but did not survive 5 the fight 6 itself 6. However, he 4 was "sent back" because its 13 task 14 is to overthrow Sauron 15, 14, had not yet met. He {4, 15} calms the three companions 1, 2, 3 of the whereabouts of the hobbits in which he {4, 15} told them 1, 2, 3 briefly how they 1, 2, 3 fared in Fangorn, and then rode with them 1, 2, 3 to Edoras to Aragorn 1, Legolas 2 and Gimli 3 can deliver on its 13 promise to Eomer. (Google ©)

In the original text there are at least 19 different relations some of which are anaphoric, while others represent non-anaphoric relations. It should be noted that we did not mark either temporal relations within the texts, or tense relations. In the translation by rule-based MT system Promt we find just 13 relations, moreover, some of the coreference chains are not correct and in some ambiguous cases (when there occurs the German pronoun « sie » which can be interpreted either as third person plural or third person singular feminine) we wrong choice has been made. In the translation by another rule-based system SDL Trados there are 15 relations. But despite a larger number of relations the overall quality is less as in most ambiguous cases the system puts a neutral pronoun « it ». Therefore, there are more shorter coreference chains as there really are. Moreover there is a dubious case of finding the right antecedent to the pronoun. In the translation by

Google there are 15 different types of relations. Yet as with the previous system there are dubious cases and short coreference chains that should have been longer.

Although quite a large number of anaphora resolution systems are available on the Web, most of them rely heavily on the quality of the syntactical analysis of the text. Some of the mistakes in anaphora resolution systems can be explained by the poor quality of the parser used in the system. However, the number of coreference chains and their length does not depend on the results of syntactical analysis. These things depend on the analysis of the input text as a whole. But there is no such a thing as a whole text analysis in any system whether it is rule-based or statistical!

In order to create a whole text analysis component we need a more thorough study of the ways a human binds sentences into a coherent text. Over the past ten years there have appeared a number of resources for anaphora resolution. Among them are large treebanks for English, German, Czech and anaphora and coreference annotated corpora (ACE corpus, ARRAU corpus, Anawiki Phrasedetectives project). But these resources are still relatively small – no more than 1 million words each. Furthermore, the number of languages for which resources are available is unfortunately insignificant[3].

If we pursue the goal of improvement the machine translation quality at a text level, then we need to make anaphora and coreference annotated parallel corpora which will provide us with comparative typological data that is required by machine translation necessities. What is more, these corpora should make use not only of nominal anaphora and coreference as is with the corpora that we have today, but also ellipsis, synonymy and other discourse relations.

Moreover, a new type of analysis should be introduced into MT, namely output text analysis. This analysis will enable the system to compare the output and the input structure, thus a higher result will be achieved.

What can NLP do for machine translation improvement?

Machine translation and its place within NLP

Machine translation occupies a special place within the field of NLP. Partially this is due to the fact that machine translation is “to blame” for the very existence of NLP and computational linguistics. For a long time MT was the main sphere of new linguistic applications. From the very beginning of MT it has been thought of as a complex linguistics system consisting of at least morphological and syntactical analysis and synthesis. For many years MT needs made researchers develop more sophisticated software for handling various linguistic aspects.

Now MT is not only a complex NLP system itself but also an essential part of many complex multilingual NLP systems. As we have mentioned earlier MT on the Web is

either a part of a web browser (a complex system of multilevel multitask analysis) or a part of a multilingual information retrieval systems. In our opinion a very essential task for MT will be merging with sentiment analysis / opinion mining systems.

Sentiment analysis and machine translation

Large amounts of data on the Web express personal subjective attitude towards different products, services, political issues, etc. For many companies, organizations and just ordinary people it is essential to know what other people think about this or that thing. But since it is impossible to read all the data it has to be processed. Such systems that process this type of data are united under the name sentiment analysis (sometimes a synonymous term – opinion mining – is used).

In the sphere of sentiment analysis research there have been successful attempts to use machine translation methods to improve the quality of the systems. The research, however, was directed at the creation of dictionaries for the needs of sentiment analysis.

On the contrary, we propose to use sentiment analysis to improve the quality of MT, by obtaining better grained data from emotionally rich sources. Moreover, emotion text representations in different languages may not correspond exactly. And that may lead to MT mistakes in the translation of this data. So in order to achieve better results in MT of opinion resources we need to further investigations on emotions, opinions and their text representations in various languages.

Conclusion

Web-based machine translation makes mistakes and it seems that it will still make them as new words and personal names appear every day. Yet it is possible to reduce the number of mistakes by using linguistically annotated parallel corpora.

As for MT in general, the most important domains for economics should be determined. And special parallel corpora have to be designed for them. The very structure of these corpora should be changed, as they should consist of various clearly defined subcorpora and not of just literature texts as we now have for some languages (The Russian National Corpus, for example).

The MT systems on the web should also make use of differentiated parsed parallel corpora. This will help to get better results for various Internet subdomain translations. Automatic domain-detection should become essential part of such systems. The work on the analysis of statistical features for special domains is already going on. We can hope that this will enable to improve web-based MT in the next 5 years.

The second task is going beyond the sentence level and dealing with whole texts. An output text analysis is an essential part of this task. And with resources becoming available we can hope for solving this task in the next 10 or 15 years.

Integration of MT with other NLP systems is another task for the future. We believe that integration with sentiment analysis will improve the MT output of personal information expressing various emotions. But integration of MT with sentiment analysis systems is not a task for the near future as much further work on emotions is necessary for such an integration,

And if machine translation is going to be so good what will the traditional translators do? It seems that they will focus on oral translation and translation of fiction, which are hard to formalize. Moreover, no MT system will be able to cope with translation of poetic texts. And as there are not so many languages with resources necessary for proper MT there is still much work for translators where they have no rivals in the near future.

Bibliography

Marchuk, Yuri (2010), « Russian language and scientific and technical translation » in *Proceedings of 4th International Congress of Russian language researchers*, Moscow, Moscow State University, p. 526. (in Russian)

Koehn, Philipp (2009), « Statistical Machine Translation », Cambridge, Cambridge University Press.

Notes

[1] Natural Language Processing

[2] German Wikipedia ©, part from the article “Der Herr der Ringe”

[3] English, Italian and small corpora (consisting of less than ten thousand words) for a handful of other European languages.