Esperus: the First Step to Build a Statistical Machine Translation System for Esperanto and Russian Languages

Darja Orlova
Saint Petersburg State University
Saint Petersburg, Russia
frenezo@mail.ru

***Abstract*—In this paper we describe an attempt to build statistical machine translation software named Esperus, which is aimed to translate from Esperanto to Russian. We apply the Moses toolkit on several corpora, consisting of novels, translated from Russian to Esperanto, and OPUS corpus. The quality of translation is evaluated with an automatic metric BLEU score, and some of the results are compared with the translation of Google Translate service. Furthermore, we analyze 400 translated sentences to find the most frequent errors and make conclusions from them.**

I. INTRODUCTION

Statistical machine translation is nowadays a cheap and easy way to build a machine translation system. It requires only parallel text corpora and a special toolkit. To improve the translation, researchers just need to enlarge the corpora or to sort the data in order to make it cleaner.

In spite of this fact many languages still cannot be translated automatically with proper quality. One of them is Esperanto. Despite being the most popular artificial language in the World it has few electronic dictionaries and even less translation systems. The most popular one - Google Translate [1] - use English as Interlingua, therefore, the result of the translation are sometimes unreadable.

In our experiment we try to build a new translation system, which will translate directly from Esperanto to Russian. Our main hypothesis is that we should obtain better results than Google Translate even with smaller parallel text corpora.

The paper is structured as follows. Section II and Section III briefly describe the process of building the statistical machine translation system. Section IV shows the evaluation of the results of our system Esperus and in Section V we infer the main problems of our project. Finally, the paper is concluded and future plans are set.

II. DATA

The modern statistical machine translation systems are phrase-based, i.e. they choose a translation according to the phrase table. To build a phrase table big enough to get an adequate translation we need a huge amount of parallel data called bitexts.

Today there are many already made open parallel corpora that can be used for research, for example, Multi-UN or Europarl. The best data for Moses toolkit [2] would be some political documents, translated in both languages, because they are written in an unambiguous and standardized language.

Unfortunately there is no government in the world, that speaks and writes in Esperanto. For that reason the number of already parallel data on the Internet is limited. That means that we needed not only to find already parallel data, but also find ways to align

A. Already parallel data

The only website that provided us some parallel data was the OPUS corpus [3] - an open-source parallel corpus, considered by some researchers as the biggest free and open corpus in the World. The information in this corpus was already collected from the Internet and can easy be downloaded in any preferred format. The data we have chosen to use is as follows:

- Text messages from KDE4, Ubuntu and GNOME software in Russian and Esperanto languages. (0,25 million target tokens).
- "OpenSubtitles2011" texts, which were previously the subtitles for different movies. (nearly 5 thousand target tokens).
- Sentences from "Tatoeba" multilingual project, which is similar to Wikipedia: anyone, who knows a language pair, can add a sentence or a translation. This corpus appeared later as a most qualitative one (0,2 million target tokens).

B. Raw data

Already collected data was not enough to build a working machine translation system. For that reason we decided to choose some novels to transform them into bitexts. Because of the copyright we were forced to find some hundred-year-old fiction novels of Russian authors, for example, Nikolaj Gogol, Ivan Turgenev and Mikhail Bulgakov. Also we added some Polish authors, because the translation to both languages seemed pretty close.

ISBN 978-952-68397-0-7 (paperback), ISBN 978-952-68397-1-4 (PDF)

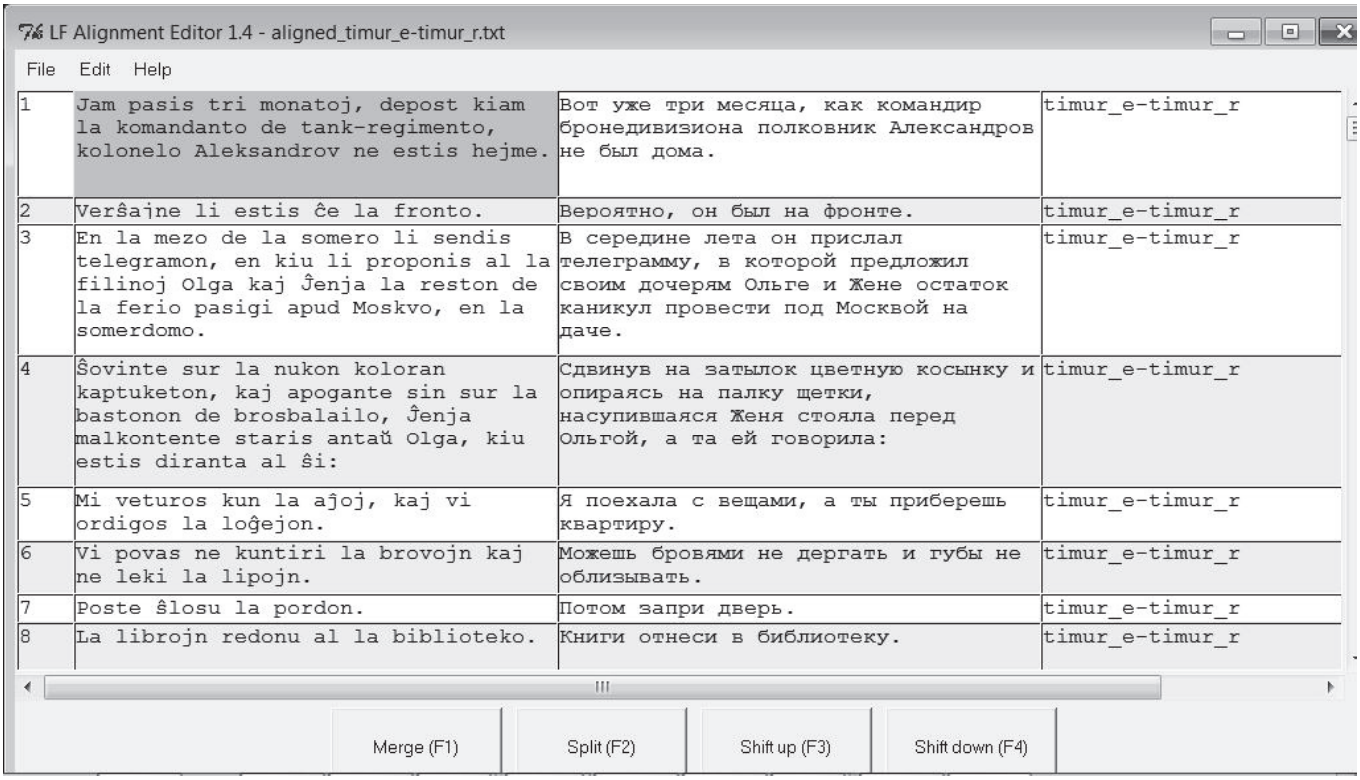

Fig. 1. LF Aligner graphical editor

However, we faced several problems with texts, originally written in English. The Russian translation sentences often had different structures due to different language features. So, we excluded some books that from the first sight seemed acceptable.

To get a bitext from 2 novels we used LF Aligner software [4]. Based on a hunalign algorithm [5], it provides decent results to other language pairs, for example, English and Russian. The program allows to choose, which pair of languages do you need and which output will you get in the end. However, when trying to match Russian and Esperanto sentences, LF Aligner somehow tried to use built-in English dictionary, and that made the output results worse. For that reason the output bitexts were always checked manually with graphical editor. (Fig. 1)

In the upshot, we collected 1.27 million source tokens and 1.35 target tokens, and that was enough to make an experiment in building a machine translation system.

III. BUILDING THE SYSTEM

The machine translation system was build using PROMT DeepHybrid Training Server (Fig.2). PROMT training server is practically a graphical interface for Moses toolkit, an open-source statistical machine translation system that allows to automatically train translation models, with some addition. Moses toolkit itself consists of GIZA++ [6], which finds parallel tokens, and the built-in implementation of model optimization (Minimum ErrorRate Training, MERT [7]). In addition, we used a readymade language model, based on publicist texts on the Internet. This language model is based on KenLM [8] libraries that are currently considered by PROMT as the best among other packages that perform language model queries.

Despite the fact that our data were more suitable for building a Russian-Esperanto translation model, not Esperanto-Russian, we decided to stick to the latter. We assumed that for us it would be easier to evaluate a Russian text, and the main disadvantages and weak points of the system would be visually more obvious.

There are features of our translation system that I would like to highlight. First of all, we transformed all the letters "ë" of Russian language into "e". The main reason for this was the fact that many modern authors ignore the former and use the latter instead. To avoid the variabily we standardized the writing manner. However, the unusual Esperanto letters ĉ, ĝ, ĥ, ĵ, ŝ, ŭ remained to avoid ambiguity.

Then, we saved in our bitexts not only tshort and simple sentences, but also compound and complex sentences, because they were often seen in the fiction texts. If we had excluded the long sentence, we would have lost an appreciable amount of data. For the same reason we left the non-punctuation symbols in sentences, because they are in common in technical texts.

Then, we excluded 1% from our training set. This 1% we used later to test our system.

In the end, we performed tuning to improve som translation model parameters. (Fig. 2) To do this we also excluded 1% of all data before the training has even started.

After taking these steps we got a machine translation system, that we called "Esperus". The name of our system, on the one hand, has abbreviations of the language pair -

Esperanto and Russian, and, on the other hand, can be translated from Esperanto as "(I) would hope", which means, that our system is only a first step to build something, but we still hope that it can grow in something more serious.

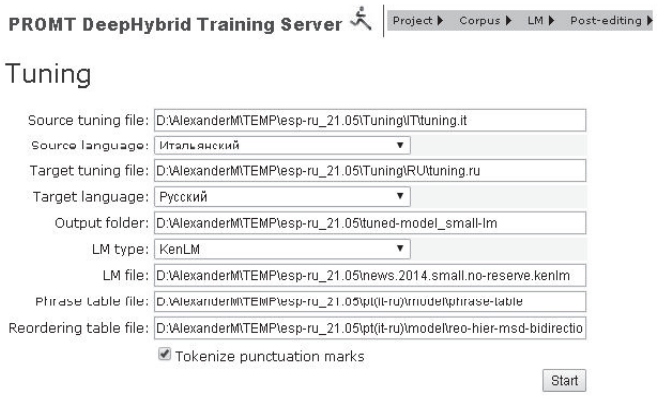

Fig. 2. PROMT server graphical interface

IV. Evaluation

We tried to evaluate the output of our system by several metrics: BLEU, comparing with Google translate and flagging the errors.

A. BLEU

BLEU Score [9] is one of the most popular methods to evaluate results of the machine translation. To use this metric, we need to compare a translation of a system with that of a human. The results may be presented on the scale from 0% to 100%, where the latter means that the human and automatic translations are identical. The BLEU score is based on comparison of n-grams. It is language independent, fast to use and unbiassed. What is more, BLEU is built in the Minimum ErrorRate Training and it helps MERT to improve translation.

To evaluate our results, we used a browser application called iBLEU [10]. It takes 3 files (original text, automatic translation, ideal translation) as an input and a BLEU Score and graphical histogram as an output. It also allows comparing several translations with the ideal one.

To evaluate our texts we divided them into 3 groups - fiction texts, that mostly consisted of novels and had long and complicated sentence structures, spoken texts, that consisted of Open Subtitles and Tatoeba corpora and had usually short simple sentences with one clause, and technical texts, taken from Ubuntu and other software.

From each group we took 1% of all sentences (70 thousand tokens totally) and ran the iBLEU application for them.

The results are shown in Table I:

TABLE I. The BLEU score for different texts

	BLEU Score
Fiction	30,82%
Spoken	26,57%
Technical	16,34%
Average	27,39%

For inflectional languages such results may be considered as successful.

V. Comparing with Google Translate.

Google Translate is one of the most powerful machine translation systems, available online. Recently Google added Esperanto to their system.

Fig. 3. A histogram comparing BLEU score for Bulgakov's text in translation of Esperus (on the top) and Google

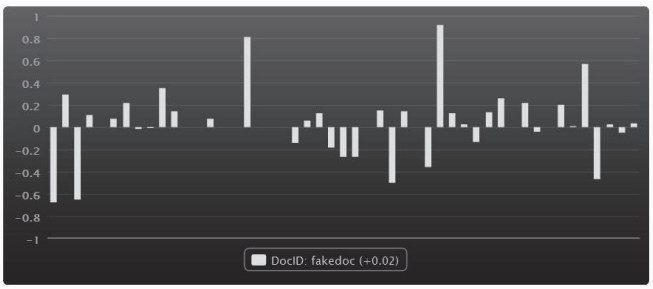

Fig. 4. A histogram comparing BLEU score for Tatoebas's text in translation of Esperus (on the top) and Google

For that reason we decided to compare our results. Thus we took 50 test sentences from the novel "The Master and Margarita" by Mikhail Bulgakov, which was used in our training set, and 50 sentences from Tatoeba corpus. We expected that the score for the former would be higher for Esperus, because it is familiar of proper names of this novel, but we could not be so sure about the latter.

The results for fiction and spoken sentences can be seen on Fig. 2 and Fig. 3 respectively.

The first group of sentences showed a pretty convincing result: 46.52% score for Esperus translation VS 15.22% for Google.

As an example we can mention this group of sentences:

- *English translation: Cry to him!*
- *Original text: Крикните ему!*
- *Esperus translation: Крикните ему!*
- *Google translation: Cry ему,*

Here we can clearly see that this mistake is caused by English language as Interlingua. Google started to translate a word from Esperanto to English, but then for some reason stopped and left it like this.

Tatoeba sentences were more complicated for Esperus, but still it managed to gain better score: 33.00% VS 31.07%.

Let us see some examples:

- *English translation: He is a Frenchman*
- *Original text: Он француз.*
- *Esperus translation: Он француженка. /He is a Frenchwoman/*
- *Google translation: Он является французский. /He appears french/*

Here we can see that Esperus probably chose the wrong translation from the table, but the Google translator made the sentence more complicated.

VI. Flagging the errors

However, there are several disadvantages of the BLEU Score metric. The main point of a machine translation is to make the output understandable to someone, who needs a translation quickly. Dealing witch languages like Russian, that has free word order and many inflections, statistical machine translation system usually fail to choose the right case or number for a noun, for instance. Being still readable enough for a user, a sentence with wrong inflections may have a low BLEU Score. The same situation we may deal with, when the translation system changes the word order, because the BLEU evaluation is simply based on n-grams.

This is the reason why many automatic translations get low scores for a Russian translation. Many researches show, that getting a result higher than 15% is almost impossible.

To evaluate our results more properly we chose the most common mistakes in machine translations and then analyzed 400 sentences, manually counting all the mistakes we see. Some sentences had several mistakes, 130 sentences had no mistakes at all. The results can be seen in Table II.

TABLE II. Number of errors in each category

	Errors
Verb tense	0
Verb number	36
Gender	39
Verb person	39
Noun/adjective case	103
Noun/adjective number	0
Wrong word	11
Wrong lexical variant	71
No translation	141
Wrong pronoun	17
Punctuation	24
Word order	0
Missing word	21
Extra word	10
Sentence structure	16
Total	528

As we can see, there were no mistakes connected with tense, adjective or noun number and word order. The translations were almost correct in these situations owing to the grammar rules of Esperanto. First of all, the word order in Esperanto is as free as in Russian, but usually it tends to be Subject-Verb-Object. Then, there are special inflections that mark the verb tense and the noun number. It means that there is only one way to translate it.

Other errors, especially connected with wrong lexical variant, can probably be solved only by building different translation systems for different text genre.

We shall mention some typical mistaken sentences:

- *No translation:* ironiis Шерсикова.
- *Wrong lexical variant:* Имеешь всегда глупую смайлик, когда надо спать.
- *Wrong gender agreement:* Это мне сказал один ворона, когда я поссорилась с ней.
- *Wrong case agreement:* Вы про опасности из-за этой фамилии или вы про опасности от люди?
- *Wrong pronoun:* Ты не понимаю это, потому что у вас нет мозги, - пламенно ответила девочка.
- *Wrong verb number agreement:* Элли встала и, пока дровосек плакал, терпеливо вытерли слезы полотенцем.

VII. Analysis

Despite the good BLEU Score, we found many mistakes in the translation of Esperus, and it cannot be just a coincidence. We analyzed thoroughly each step that we performed and found several disadvantages in what we had done.

The one and the biggest disadvantage is lack of qualitative data.

On the one hand, we used some corpora from OPUS project without checking the corpora. On closer examination we saw, that some sentence in technical part of OPUS corpora are even not translated from English. It contains sometimes random symbols in the middle of the sentence, and these moments can certainly ruin all the statistics. So, we decided to exclude technical corpora from our future research.

On the other hand, we used all fiction novels that could be found on the Internet, legal and free, and it is not enough. So, now we see a perspective in Tatoeba project and Wikipedia texts. That needs that we need to encourage people around the world, who know Esperanto, to add translations to these two Websites. It would certainly help not only our project, but also other people who try to learn and improve Esperanto around the world.

The second great disadvantage of our current translator is different spelling of words with diacritic marks in Esperanto. Unfortunately, we realized that there are several ways to write words with diacritics:

1) The right way, which was created by Zamenhof: ĉ, ĝ, ĥ, ĵ, ŝ, ŭ;

2) The old-fashioned printing way, which was used by publishers in XX century. Some of them could print the diacritic mark, so they used "h" instead: ch, gh, hh, jh, sh, uh;

3) The modern "internet" way, which is used today in World Wide Web by those, who cannot install special software to type these words: cx, gx, hx, jx, sx, ux. This way is the most appreciated, because there is no "x" in Esperanto, so you can always understand that it replaces a diacritic mark;

4) The modern "lazy" way, when someone just ignores the fact of their existence: c, g, h, j, s, u

We didn't pay much attention to these, and it resulted in a variants in the phrase table and inability to translate some simple words, that are written in, for example, modern "internet" way.

To solve this problem, we need to build some spellchecker in our translator, which will unify the ways of writing into one.

The third disadvantage of our project is a problem with aligner. As we mentioned before, LF Aligner had no Esperanto dictionary inside, so it tried to match the sentences with an English one. We failed to add an Esperanto dictionary, so we were forced to check the results by ourselves. To continue the work more comfortable, we need first of all to fix this little bug, and then the precision of automatic alignment will be higher.

We hope that with dealing with these problems we will manage to improve our results in the nearest future.

VIII. Further plans

We plan to continue working on the Esperus project.

Our next step is to build a translation model from Russian to Esperanto. To do this, we will need a certain amount of monolingual data in Esperanto to build a language model. For this we plan to use Wikipedia.

Our next step is to design a graphical interface to our program, either in browser or as separate software. This interface will accept different ways of writing the diacritical symbols.

IX. Conclusion

We have succeeded in our goal to build a working prototype of a statistical machine translation system from Esperanto to Russian language that all in all gets better results than Google Translator.

We have discovered several problems underlying the incorrect translation, and some of then we still do not know how to solve. For example, the problem of choosing between "вы" (formal "you") and "ты" (informal "you") without a context.

However, in the nearest future we are able to improve our result and set new goals. We realized that the main issue of our project is lack of big and qualitative amount of parallel data, and that is usually a problem of many other statistical machine translation systems. What is more, we found some specific problems, for example, different ways of writing the diacritical marks.

We hope that our work will inspire other scientists to consider Esperanto as interest language to work with.

Acknowledgment

We would like to thank Alexander Molchanov for providing access to PROMT DeepHybrid Training Server and useful links. Also, we would like to thank Olga Mitrenina for guidance and recommended literature. In the end, we would like to thank Vjacheslav Ivanov for some ideas about parallel data and Esperanto literature.

References

[1] Google Translate official website Web: https://translate.google.com/
[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst: Moses: open source Toolkit for statistical machine translation. ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic,2007, 177–180.
[3] Jörg Tiedemann, Lars Nygaard: The OPUS corpus - parallel & free. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, 2004.
[4] LF Aligner Web Site, Web: http://sourceforge.net/projects/aligner/
[5] Hunalign User's Guide Web: http://mokk.bme.hu/resources/hunalign/
[6] Franz Josef Och, Hermann Ney: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1):19–51, Cambridge, MA, US, 2003.
[7] Franz Josef Och: Minimum Error Rate Training in Statistical Machine Translation. ACL, Stroudsburg, PA, USA, 2003
[8] Kenneth Heafield: KenLM: Faster and Smaller Language Model Queries. WMT at EMNLP, Edinburgh, Scotland, United Kingdom, 2011
[9] K. Papineni, S. Roukos, T. Ward: BLEU: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), 2001.
[10] Nitin Madnani: iBLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems. ICSC, Palo Alto, CA, US,. 2011.“A mathematical theory of communication”, *Bell Syst. Tech. J.*, vol. 27, 1948, pp. 379-423.