Building a Web-based parallel corpus and filtering out machine-translated text

Alexandra Antonova, Alexey Misyurev
Yandex
16, Leo Tolstoy St., Moscow, Russia
{antonova, misyurev}@yandex-team.ru

Abstract

We describe a set of techniques that have been developed while collecting parallel texts for Russian-English language pair and building a corpus of parallel sentences for training a statistical machine translation system. We discuss issues of verifying potential parallel texts and filtering out automatically translated documents. Finally we evaluate the quality of the 1-million-sentence corpus which we believe may be a useful resource for machine translation research.

1 Introduction

The Russian-English language pair is rarely used in statistical machine translation research, because the number of freely available bilingual corpora for Russian-English language pair is very small compared to European languages. Available bilingual corpora[1] often belong to a specific genre (software documentation, subtitles) and require additional processing for conversion to a common format. At the same time many Russian websites contain pages translated to or from English. Originals or translations of these documents can also be found in the Internet. By our preliminary estimates these bilingual documents may yield more than 100 million unique parallel sentences while it is still a difficult task to find and extract them.

The task of unrestricted search of parallel documents all over the Web including content-based search is seldom addressed by researchers. At the same time the properties of the set of potential parallel texts found in that way are not well investigated. Building a parallel corpus of high quality from that kind of raw data is not straightforward because of low initial precision, frequent embedding of nonparallel fragments in parallel texts, and low-quality parallel texts. In this paper we address the tasks of verification of parallel documents, extraction of the best parallel fragments and filtering out automatically translated texts.

Mining parallel texts from a big document collection usually involves three phases:

- Detecting a set of potential parallel document pairs with fast but low-precision algorithms
- Pairwise verification procedure
- Further filtering of unwanted texts, e.g. automatically translated texts

Finding potential parallel texts in a collection of web documents is a challenging task that does not yet have a universal solution. There exist methods based on the analysis of meta-information (Ma and Liberman, 1999; Resnik, 2003; Mohler and Mihalcea, 2008, Nadeau and Foster 2004), such as URL similarity, HTML markup, publication date and time. More complicated methods are aimed at

[1] e.g. http://opus.lingfil.uu.se/

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 136–144,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011. ©*2011 Association for Computational Linguistics*

detecting potential parallel texts by their content. In this case mining of parallel documents in the Internet can be regarded as the task of near-duplicate detection (Uszkoreit et al., 2010). All of the above mentioned approaches are useful as each of them is able to provide some document pairs that are not found by other methods.

In our experiments, fast algorithms of the first phase classify every pair of documents as parallel with very low precision, from 20% to 0.001%. That results in a huge set of candidate pairs of documents, for which we must decide if they are actually parallel or not. For example, if we need to get 100 000 really parallel documents we should check from 500 thousand to 100 million pairs. The large number of pairwise comparisons to be made implies that the verification procedure must be fast and scalable. Our approach is based on a sentence-alignment algorithm similar to (Brown et al., 1991; Gale and Church, 1993; Chen, 1993; Moore 2002; Ma, 2006) but it is mainly aimed at achieving high precision rather than high recall. The algorithm is able to extract parallel fragments from comparable documents, as web documents often are not exactly parallel. The similarity estimate relies on probabilistic dictionary trained on initial parallel corpus and may improve when the corpus grows.

Due to growing popularity of machine translation systems, Russian websites are being increasingly filled with texts that are translated automatically. According to selective manual annotation the share of machine translation among the texts that have passed the verification procedure is 25-35%. Machine-translated sentences often demonstrate better word correspondence than human-translated sentences and are easier to align, but the longer phrases extracted from them are likely to be unnatural and may confuse the statistical translation system at the training stage. The large share of automatically translated data decreases the value of the corpus, especially if it is intended for research. Also it will make it difficult to outperform the translation quality of the system which generated those sentences.

To the best of our knowledge, there is no existing research concerning the task of filtering out machine translation. Our filtering method is based on a special decoding algorithm that translates sentence-aligned document and then scores the output against the reference document with BLEU metric. This method allows reducing the number of automatically translated texts to 5% in the final corpus.

Our final goal is to build a quality corpus of parallel sentences appropriate for training a statistical machine translation system. We evaluate the 1-million-sentence part of our corpus by training a phrase-based translation system (Koehn et al., 2007) on these sentences and compare the results with the results of training on noisy data, containing automatically translated texts as its part.

The rest of the paper is organized as follows: Section 2 provides an overview of the system architecture and addresses specific problems at the preparatory stage. Section 3 describes the sentence-alignment algorithm and the pairwise verification procedure. The algorithm makes use of statistical dictionaries trained beforehand. In Section 4 we discuss the problem of filtering out automatically translated texts. In Section 5 we evaluate the quality of the final parallel corpus and provide some statistical information about Russian-English language pair. We conclude in Section 6 with short summary remarks.

2 System description

The corpus building procedure includes several stages represented in Figure 1. Initial training provides bilingual probabilistic dictionaries which are used in sentence alignment and verification of potential parallel texts. We used Russian/English correspondent pages from a number of bilingual web-sites of good quality. We performed robust alignment based on sentence lengths as in (Gale and Church, 1993). The obtained probabilistic dictionaries were gradually improved in a sort of a bootstrapping procedure when the corpus size increased.

Our main source of Web documents are web pages from search engine database with their textual contents already extracted and sentence boundaries detected. Nevertheless documents often include sentences that are site-specific and carry some meta-information, advertising, or just some noise. When often repeated such sentences may confuse statistical training, so we choose to delete subsequent sentences that have been encountered recently.

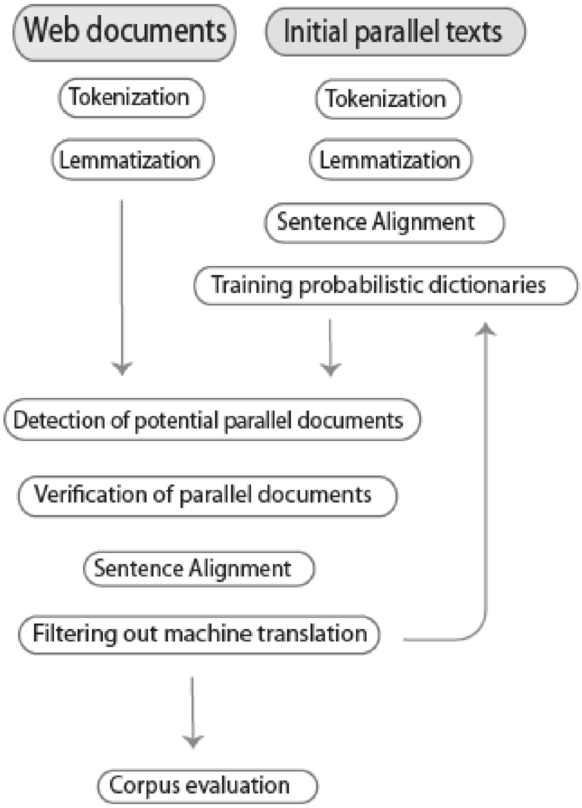
Web documents
Initial parallel texts
Tokenization
Tokenization
Lemmatization
Lemmatization
Sentence Alignment
Training probabilistic dictionaries
Detection of potential parallel documents
Verification of parallel documents
Sentence Alignment
Filtering out machine translation
Corpus evaluation

Figure 1. Corpus building procedure.

In morphologically rich languages nouns, verbs and adjectives have many different forms in text, which complicates statistical training, especially when the initial collection is comparatively small. At the same time, the task of sentence alignment relies on robust algorithms which allow for some data simplification. Word stemming, truncation of word endings and lemmatization may be used to reduce the data sparseness problem when dealing with morphologically rich languages. The accurate lemmatization algorithms for Russian language are complicated and comparatively slow because they should resolve morphological ambiguity as many word forms have more than one possible lemma. We chose a simple and fast algorithm of probabilistic lemmatization where a word is always assigned the most frequent of its possible lemmas. There are several reasons why it is appropriate for the task of sentence and word alignment:

- The algorithm runs almost as fast as the word truncation method, and in most cases it yields correct lemmas.
- Most of the information is contained in low-frequency words and those are usually less ambiguous than the frequent words.
- Individual mistakes in lemmatization do not necessarily result in wrong similarity estimation for the whole sentence.

3 Verification of potential parallel documents

Potential parallel documents are a pair of texts; each of them represents the textual content of some HTML page. The size of texts may vary from several sentences to several thousand sentences.

Our approach to the task of verification of potential parallel documents is motivated by the properties of the set of potential parallel texts, which is the output of different search algorithms including unrestricted content-based search over the Web.

The first problem is that most of the potential parallel texts on the Web, even if they prove to have parallel fragments, often contain non-parallel fragments as well, especially at the beginning or at the end. Since the parallel fragment can be located anywhere in the document pair, the verification algorithm performs exhaustive dynamic programming search within the entire document and not only within a fixed width band around the main diagonal. Our similarity measure relies heavily on features derived from the sentence alignment of the best parallel fragment and does not utilize any information from the rest of the text. We allow that the parallel fragment begins and ends anywhere in the text and also it is possible to skip one or several sentences without breaking the fragment.

We have also considered the possibility that documents can contain more than one parallel fragment separated by greater non-parallel fragments. Though such documents do exist, the contribution of lesser parallel fragments to parallel corpus is insignificant compared to much simpler case where each pair of documents can contain only one parallel fragment.

The second problem of the input data is low initial precision of potential parallel texts and the fact that there are many comparable but not parallel texts. It is worth noting that the marginal and joint probabilities of words and phrases in the set of documents with similar content may differ substantially from the probabilities obtained from the parallel corpus of random documents. For this reason we cannot completely rely on statistical models trained on the initial parallel corpus. It is important to have a similarity measure that allows for additional adjustment in order to take into account the probability distributions in the potential parallel texts found by different search algorithms.

The third problem is the large number of pairwise comparisons to be made. It requires that the verification procedure must be fast and scalable. Due to the fact that the system uses precomputed probabilistic dictionaries, each pair of documents can be processed independently and this stage fits well into the MapReduce framework (Dean and Ghemawat, 2004). For example, verification of 40 million pairs of potential parallel texts took only 35 minutes on our 250-node cluster.

The algorithm of verifying potential parallel documents takes two texts as input and tries to find the best parallel fragment, if there is any, by applying a dynamic programming search of sentence alignment. We use sentence-alignment algorithm for handling four tasks:

- Search of parallel fragments in pairs
- Verification of parallel document pairs
- Search of per-sentence alignment
- Filtering out sentences that are not completely parallel

Each sentence pair is scored using a similarity measure that makes use of two sources of prior statistical information:

- Probabilistic phrase dictionary, consisting of phrases up to two words
- Empirical distribution of lengths of Russian/English parallel sentences

Both have been obtained using initial parallel corpus. In a sort of bootstrapping procedure one can recalculate that prior statistical information as soon as a bigger parallel corpus is collected and then realign the input texts.

The algorithm neither attempts to find a word alignment between two sentences, nor it tries to translate the sentence as in (Uszkoreit et al., 2010). Instead, it takes account of all phrases from probabilistic dictionary that are applicable to a given pair of sentences disregarding position in the sentence or phrase intersection. Our probabilistic dictionary consists of 70’000 phrase translations of 1 or 2 words.

Let S and T be the set of source/target parts of phrases from a probabilistic dictionary, and $E \subset S \times T$ - the set of ordered pairs, representing the source-target dictionary entries (s,t). Let the source sentence contain phrases $S_0 \subset S$ and the target sentence contain phrases $T_0 \subset T$. Then the similarity between the two sentences is estimated by taking the following factors into account:

- $p(s \mid t)$, $p(t \mid s)$, translation probabilities;
- len_S, len_T , length of source and target sentences;
- $\log \hat{p}(len_S, len_T)$, the empirical distribution of length correspondence between source and target sentences.

The factors are log-linearly combined and the factor weights are tuned on the small development set containing 700 documents. We choose the weights so that the result of comparison of nonparallel sentences is usually negative. As a result of the search procedure we choose a parallel fragment with the biggest score. If that score is above a certain threshold the parallel fragment is extracted, otherwise the whole document is considered to be nonparallel.

Relative sentence order is usually preserved in parallel texts, though some local transformations may have been introduced by the translator, such as sentence splitting, merge or swap. Though sentence-alignment programs usually try to detect some of those transformations, we decided to ignore them for several reasons:

- Split sentences are not well suited to train a phrase-based translation system.
- One part of a split sentence can still be aligned with its whole translation as one-to-one correspondence.
- Cases of sentence swap are too rare to justify efforts needed to detect them.

4 Filtering out machine translation

After the verification procedure and sentence-alignment procedure our collection consists of sentence-aligned parallel fragments extracted from initial documents. A closer look at the parallel fragments reveals that some texts contain mistakes typically made by machine translation systems. It is undesirable to include such documents into the corpus, because a phrase-based translation system trained on this corpus may learn a great deal of badly constructed phrases.

The output of a rule-based system can be recognized without even considering its source text, as having no statistical information to rely on, the rule-based systems tend to choose the safest way of saying something, which leads to uncommonly frequent use of specific words and phrases. The differences in n-gram distributions can be captured by comparing the probabilities given by two language models: one trained on a collection of the outputs of a rule-based system and the other – on normal texts.

Our method of filtering out statistical machine translation is based on the similarity of algorithms of building phrase tables in the existing SMT systems. Those systems also have restrictions on reordering of words. Therefore their output is different from human translation, and this difference can be measured and serve as an indicator of a machine translated text. We designed a special version of phrase-based decoding algorithm whose goal was not just translate, but to provide a translation as close to the reference as possible while following the principles of phrase-based translation. The program takes two sentence-aligned documents as an input. Prior to translating each sentence, a special language model is built consisting of n-grams from the reference sentence. That model serves as a sort of soft constraint on the result of translation. The decoder output is scored against reference translation with the BLEU metric (Papineni et al., 2002) - we shall call it r-bleu for the rest of this section. The idea is that the higher is r-bleu, the more likely the reference is statistical translation itself.

The program was implemented based on the decoder of the statistical phrase-based translation system. The phrase table and the factor weights were not modified. Phrase reordering was not allowed. The phrase table contained 13 million phrases. The language model was modified in the following way. We considered only n-grams no longer than 4 words and only those that could be found in the reference sentence. The language model score for each n-gram depended only on its length.

We evaluated the method efficiency as follows. A collection of 245 random parallel fragments has been manually annotated as human or machine translation.

There are some kinds of typical mistakes indicating that the text is generated by a machine translation system. The most indicative mistake is wrong lexical choice, which can be easily recognized by a human annotator. Additional evidence are cases of incorrect agreement or unnatural word order. We considered only fragments containing more than 4 parallel sentences, because it was hard to identify the origin of shorter fragments. The annotation provided following results:

- 150 documents - human translation (64% of sentences)
- 55 documents - English-Russian machine translation (22% of sentences)
- 32 documents - Russian-English machine translation (12% of sentences)
- 8 documents - not classified (2% of sentences)

Sometimes it was possible for a human annotator to tell if a translation has been made by a rule-based or phrase-based translation system, but generally it was difficult to identify reliably the origin of a machine translated text. Also there were a number of automatically translated texts which had been post-edited by humans. Such texts often preserved unnatural word order and in that case they were annotated as automatically translated.

The annotation quality was verified by cross-validation. We took 27 random documents out of 245 and compared the results of the annotation with those performed by another annotator. There was no disagreement in identifying the translation direction. There were 4 cases of disagreement in identifying automatic translation: 3 cases of post-edited machine translation and 1 case of verbatim human translation. We realized that in case of post-edited machine translation the annotation was subjective. Nevertheless, after the question was discussed we decided that the initial annotation was correct. Table 1 represents the results of the annotation along with the range of r-bleu score.

r-bleu	Human	Automatic
0 - 5	0	0
5-10	252	0
10-15	899	0
15-20	1653	0
20-25	1762	0
25-30	1942	154
30-35	1387	538
35-40	494	963
40-45	65	1311
45-50	76	871
50-55	23	658
55-60	0	73
Total	8553	4568

Table 1. Number of parallel sentences in human/machine translated documents depending on the range of r-bleu score.

Let $C_{h\max}$ denote the total number of sentences in all documents which were annotated as human translation. In our case $C_{h\max} = 8553$. Let C_h denote the number of sentences in human translated documents with a r-bleu beyond certain threshold, and C_{mt} – the number of sentences in automatically translated documents with a r-bleu beyond the same threshold. Then recall(R) and precision(P) are defined as

$$R = C_h / C_{h\max} ,$$
$$P = C_h / (C_h + C_{mt}).$$

For example, if we discard documents with r-bleu > 33.0, we get R = 90.1, P = 94.1. Figure 2 illustrates the dependency between these parameters.

The evaluation showed that parallel documents that have been translated automatically tend to get higher r-bleu scores and may be filtered out with reasonable precision and recall. As it is shown in Table 1, the total rate of machine translated sentence pairs is about 35% before the filtration. According to manual evaluation (see section 5, Table 4), this rate is reduced down to 5% in the final corpus.

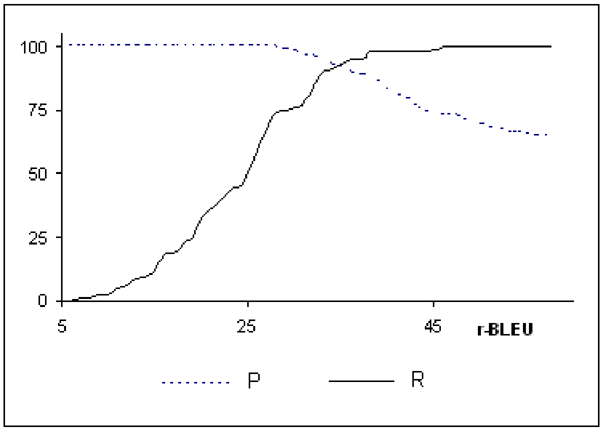

Figure 2. Dependency between r-bleu score and recall(R)/precision(P) rates of filtering procedure.

We chose the BLEU criterion partly due to its robustness. For the English-Russian language pair it yielded satisfactory results. We believe that our approach is applicable to many other language pairs as well, probably except the pairs of languages with similar word order. For those languages some other metric is possibly needed taking into account properties of particular language pair. We expect that the r-bleu threshold also depends on the language pair and has to be re-estimated.

5 Corpus of parallel sentences

After we choose a threshold value of the r-bleu criterion, we remove texts with the r-bleu score higher than the threshold from our collection of parallel fragments. Then we extract parallel sentences from the remaining texts in order to get a corpus of parallel sentences.

Sentences inside parallel fragments undergo some additional filtering before they can be included into the final corpus. We discard sentence pairs for which a similarity score is below a given threshold, or word-length ratio is less than ½. It is also useful to drop sentences whose English part contains Cyrillic symbols as those are extremely unlikely to be seen in original English texts and their presence usually means that the text is a result of machine translation or some sort of spam. All sentence pairs are lowercase and distinct. Sentences of more than 100 words have been excluded from the corpus.

In the rest of this section we estimate the quality of a 1-million-sentence part of the final parallel corpus that we are going to share with the research community. The corpus characteristics are represented in Table 2 and examples of parallel sentences are given in Table 3.

	English	Russian
Sentences	1`022`201	
Distinct sentences	1`016`580	1`013`426
Words	27`158`657	25`135`237
Distinct words	323`310	651`212
Av. Sent. Len	26.5	24.6

Table 2. Corpus characteristics: number of parallel sentences, distinct sentences, words[2], distinct words and average sentence length in words.

We evaluate corpus quality in two ways:

- Selecting each 5000-th sentence pair from the corpus and manually annotating the sentences as parallel or not. The results of the manual annotation are represented in Table 4.
- Training a statistical machine translation system on the corpus and testing its output with BLEU metric

We trained two phrase-based translation systems[3]. The first system was trained on 1 million random sentences originated in the documents which were human translations according to our r-bleu criterion. The other system was trained on the same corpus except that 35% of sentences were replaced to random sentences taken from documents which had been previously excluded as automatically translated. We reserved each 1000-th sentence from the first “clean” corpus as test data. We get word-alignment by running Giza++ (Och et al., 2000) on lemmatized texts. The phrase-table training procedure and decoder are the parts of Moses statistical machine translation system (Koehn et al., 2007). The language model has been trained on target side of the first corpus using SRI Language Modeling Toolkit (Stolcke, 2002).

в 2004 майдан прославился на весь мир благодаря оранжевой революции, которая происходила на этой площади. in 2004 maidan became-famous over all world due-to orange revolution , which took-place at this place . in 2004, maidan became famous all over the world because the orange revolution was centered here.
рассказы о народах, чей язык настолько несовершенен, что он должен восполняться жестами, - чистые мифы. stories about peoples , whose language so-much imperfect , that it should be-supplied gestures-with , - pure myths . tales about peoples whose language is so defective that it has to be eked out by gesture, are pure myths.
остальное время пусть они будут открыты, чтобы все обитатели вселенной могли увидеть тебя! the-rest-of time let they be open , so-that all inhabitants universe-of could see you ! the rest of the time, let the doors be open so that all the residents of the universe may have access to see you.
"я контролирую свою судьбу. "i control my destiny. "i control my own destiny.

Table 3. Sample parallel sentences.

Parallel	169
Parallel including non-parallel fragments	19
Non-parallel	6
English-Russian automatic [4] translation	7
Russian-English automatic translation	3
Total sentences	**204**

Table 4. Results of manual annotation of 204 sample sentences from the corpus.

[2] Punctuation symbols are considered as separate words.

[3] http://www.statmt.org/moses/

[4] Sentences containing mistakes typical for MT systems were annotated as automatic translations.

We tested both Russian-to-English and English-to-Russian translation systems on 1022 test sentences varying the language model order from trigram to 5-gram. We have not tuned the weights on the development set of sentences, because we believe that in this case the quality of translation would depend on the degree of similarity between the test and development sets of sentences and it would make our evaluation less reliable. In all experiments we used default Moses parameters, except that the maximum reordering parameter was reduced to 3 instead of 6. The results are represented in Table 5.

	Ru-En / +mt	En-Ru / +mt
3-gram	20.97 / +0.06	16.35 / -0.10
4-gram	21.04 / -0.13	16.33 / -0.13
5-gram	21.17 / -0.06	16.42 / -0.16
OnlineA[5]	25.38	21.01
OnlineB[6]	23.86	16.56

Table 5. BLEU scores measured on 1022 test sentences depending on the order of language model. The column +mt shows relative change in BLEU score of the system trained on "mt-noisy" data.

The overall system performance can be improved by tuning and/or training a bigger language model, but our goal is only to show to what extent the corpus itself is suitable for training statistical machine translation system. Online translation systems have been tested on the same test set, except that the input was detokenized and the output was lowercased. The online translation could have been better if the input text was in its original format - not lowercased.

6 Conclusion

We have described our approaches to main problems faced when building a parallel Russian-English corpus from the Internet.

We have proposed a method of filtering out automatically translated texts. It allowed us to reduce the rate of sentence pairs that originate from machine translated documents from 35% to 5%. The approach relies on general properties of the state-of-the-art statistical translation systems and therefore is applicable to many other language pairs.

We presented results of evaluation of the resulting Russian-English parallel corpus. We believe that the 1-million-sentence Russian-English corpus of parallel sentences used in this paper is a useful resource for machine translation research and machine translation contests.

References

Brown, P.F., Lai, J.C., Mercer, R.L. 1991. Aligning Sentences in Parallel Corpora. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California 169–176.

Chen, S.F. 1993. Aligning sentences in bilingual corpora using lexical information. Conference of the Association for Computational Linguistics, Columbus, Ohio, 9-16.

Dean, J. and Ghemawat, S. 2004. MapReduce: Simplified data processing on large clusters. In Proceedings of the Sixth Symposium on Operating System Design and Implementation (San Francisco, CA, Dec. 6–8). Usenix Association.

Gale, W. A., & Church, K. W. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(3), 75-102.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June.

Xiaoyi Ma and Mark Liberman. 1999. BITS: A method for bilingual text search over the web. Proceedings of the Machine Translation Summit VII.

Xiaoyi Ma. 2006. Champollion: A Robust Parallel Text Sentence Aligner. LREC 2006: Fifth International Conference on Language Resources and Evaluation.

Michael Mohler and Rada Mihalcea. 2008. BABYLON Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages. Proceedings of the Language Resources and Evaluation Conference.

Moore, Robert C., 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. Machine Translation: From Research to Real Users

[5] http://translate.google.ru/

[6] http://www.microsofttranslator.com/

(Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, 135-244.

David Nadeau and George Foster, 2004. Real-time identification of parallel texts from bilingual newsfeed. Computational Linguistic in the North-East (CLiNE 2004): 21-28.

Franz Josef Och, Hermann Ney. 2000. Improved Statistical Alignment Models. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318, Philadelphia, PA, USA.

Resnik, Philip and Noah A. Smith. 2003. The web as a parallel corpus. Computational Linguistics, 29:349–380.

Andreas Stolcke. 2002. SRILM—an extensible language modeling toolkit. Proceedings ICSLP, vol. 2, pp. 901–904, Denver, Sep.

Jakob Uszkoreit, Jay Ponte, Ashok Popat and Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. Coling