Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Automatic Bilingual Phrase Dictionary Construction from GIZA++ Output

Albina Khusainova
Innopolis University
Innopolis, Russia
a.khusainova@innopolis.ru

Vitaly Romanov
Innopolis University
Innopolis, Russia
v.romanov@innopolis.ru

Adil Khan
Innopolis University
Innopolis, Russia
a.khan@innopolis.ru

Abstract

Modern encoder-decoder based neural machine translation (NMT) models are normally trained on parallel sentences. Hence, they give best results when translating full sentences rather than sentence parts. Thereby, the task of translating commonly used phrases, which often arises for language learners, is not addressed by NMT models. While for high-resourced language pairs human-built phrase dictionaries exist, less-resourced pairs do not have them. We suggest an approach for building such dictionary automatically based on the GIZA++ output and show that it works significantly better than translating phrases with a sentences-trained NMT system.

Keywords: phrase translation, machine translation, automatic bilingual dictionary construction, phrase dictionary, language resources

DOI: 10.28995/2075-7182-2022-21-1068-1077

Автоматическое построение билингвального словаря на основе вывода GIZA++

Альбина Хусаинова
Университет Иннополис
Иннополис, Россия
a.khusainova@innopolis.ru

Виталий Романов
Университет Иннополис
Иннополис, Россия
v.romanov@innopolis.ru

Адил Хан
Университет Иннополис
Иннополис, Россия
a.khan@innopolis.ru

Аннотация

Современные модели нейронного машинного перевода (НМП) на основе энкодера-декодера, как правило, обучают на корпусах параллельных предложений. Соответственно, такие модели выдают наилучшие результаты при переводе полных предложений, а не их частей. Таким образом, подобные модели не решают задачи перевода устойчивых выражений, которая часто возникает при изучении языка. И если для высокоресурсных языковых пар бывают доступны словари фраз и выражений созданные вручную, для более низкоресурсных пар их чаще всего просто не существует. В этой работе мы предлагаем автоматический подход к созданию такого словаря на основе вывода GIZA++ и демонстрируем, что он работает значительно лучше, чем перевод фраз с помощью системы НМП, обученной на предложениях.

Ключевые слова: перевод фраз, машинный перевод, автоматическое построение билингвального словаря, языковые ресурсы

1 Introduction

Second language learners and users typically utilize their first language to find translations to the words, phrases, and sentences in a second language. People translate sentences when there is a ready text in a source language, whether it is copied or composed by a user. However, when a user forms a sentence right away in a second language, s/he often needs to consult a dictionary for the correct translation of a word or a phrase, and this is especially true for writing in the second language (Jun, 2008).

Learning the vocabulary of words in the second language is a basic step. However, it is not enough to know individual words, since most of the time it is phrases that play the role of semantic units, not words, so studying collocations is essential (Vasiljevic, 2014). For this reason, good language learning tools always teach words and phrases together, so that the user is able to understand and form coherent sentences based on them. Thus, for creating language learning tools it is necessary to have not only word dictionaries but also high-quality phrase dictionaries.

For second language users, on the other hand, the need for phrase dictionaries also arises in many contexts. For instance, when reading texts that contain unfamiliar words or phrases—a good example is e-books that have tooltips with dictionary items. Users might be interested in a phrase translation directly or, if they come across a new word, they might want to know the common collocations of that word together with their translations, which also leads to phrase dictionaries.

Another common use case is writing in a second language. When the idea is being verbalized, a user either immediately recalls the needed words and collocations or, otherwise, has to first translate them from the first language. In the latter case, it is very important to provide the user with a list of possible translations such that s/he can choose the one that carries the intended meaning and best matches the context. Providing such lists is only possible if corresponding language resources (dictionaries) exist.

Word-level translations can usually be found in human-built dictionaries, and sentence translations can typically be obtained using online NMT tools. However, when it comes to phrases, the situation is different. Usually, only rich-resourced language pairs do have good manually constructed bilingual common phrase dictionaries. Still, they are often incomplete, or too narrow, for example, limited to noun phrases. As for the neural translation, models trained on whole sentences often do not provide high-quality output for phrases—it can be simply erroneous or there can be a single translation while actually there exist many equally good alternatives. This is frequently alleviated by incorporating data from existing dictionaries—when a user searches for a common phrase translation, the system switches from neural translation to simple dictionary lookup. However, as already mentioned, such dictionaries often do not exist for many language pairs.

In this work, we suggest a way to construct a bilingual phrase dictionary automatically based on a corpus of parallel texts. We retrieve candidate translations from a phrase table which are built based on the output of the statistical tool GIZA++ (Brown et al., 1993; Och and Ney, 2003) and then filter and sort them using heuristics. As a result, we get a phrase dictionary that can be used as-is or can serve as a basis for a manually constructed dictionary. We examine the resulting dictionary and measure its quality against the golden standard and NMT translation. Finally, we make the constructed Russian-English phrase dictionary available online as a linguistic resource.

2 Related work

Phrase translation as a separate task is not presented in the literature. However, there are some, mostly older, works on *collocation translation*. Since the term *collocation* is very related to the term *phrase* as we understand it, we consider the literature on collocation translation to be relevant. The most recent work (Garcia et al., 2019) suggests using word embeddings to find bilingual collocations—first mapping collocation *bases* and then their possible collocates. The limitation of such approach is that it restricts collocation translations to very exact correspondences only, whereas quite often phrases can be more idiomatic. Also, according to their approach, the number of words in a collocation should correspond to the number of words in its translation, which is also often not the case. For example, English phrase 'bring about' can be translated as a single word 'вызывать' (vyzyvat') in Russian.

As for earlier works, Smadja et al. (1996) translate collocations word by word by maximizing Dice coefficient scores between source and target collocations in a parallel corpus. They make an assumption that any source collocation has a unique translation in the target language, which is not very realistic. In a similar manner, Kupiec (1993) separately extracts noun phrases in two languages and maximizes their co-occurrence using a bilingual corpus.

Rivera et al. (2013) assume that collocations in both languages have the same part of speech (POS) structure. Using dictionaries, they find a translation for a *base* word and then search for co-occurring target language collocations with the same POS-structure in the sentences of a parallel corpus. Seretan and Wehrli (2007) employ a similar approach where bilingual dictionaries are used to find *base* translations and syntactic parsing is applied to find corresponding collocations.

In our case, phrases are not in general expected to have the same syntactic or POS-structure. Also, since we do not focus on collocations only, choosing the *base* word might be ambiguous. Hence, we do not consider approaches that match *base* words and rely on syntactic/POS correspondences.

Instead, we are inclined towards methods that find phrase translations using word alignment. One of the strongest statistical tools for aligning words in parallel sentences is GIZA++ (Brown et al., 1993; Och and Ney, 2003). Although the underlying IBM word alignment models were developed decades ago, GIZA++ still cannot be fully outperformed by modern neural methods. Only recently some works (Zenkel et al., 2020; Chen et al., 2020b) which employ neural architectures were able to show some improvements over GIZA++. However, the analysis shows that these improvements are due to better recall but not precision. In our case, precision is more important, since when constructing a dictionary, it is better to have fewer but more accurate results.

When the words are aligned in both source-to-target and target-to-source directions, the resulting alignments are combined using the 'grow-diag' method (Koehn et al., 2005). The phrases are then extracted and aligned based on the *consistency* criteria: "The words in the phrase pair have to be aligned to each other and not to any words outside" (Koehn et al., 2005). As a result, there is a list of phrases with their possible translations, scored by their probabilities. It is called a *phrase table* and it was originally intended to be a part of the statistical machine translation system. Nowadays, statistical machine translation is replaced by neural machine translation, however, this by-product, a phrase table, still proves to be useful.

Works similar to ours which use phrase tables to build/extend bilingual dictionaries include Richardson et al. (2014), Daiga Deksne (2018), and Chen et al. (2020a). The next section describes our approach in full detail.

3 Methodology

We aim at constructing a phrase dictionary, and we need to define what we mean by *phrase*. We understand phrase as an n-gram of words that carry some clear meaning, co-occur more often than simply by chance (as collocations), and whose overall meaning may not necessarily be understood from the individual words (as idioms). We need to note that due to the chosen alignment method's restriction, we only consider contiguous phrases.

Usually, when constructing a bilingual dictionary, the first step is to identify the collocations/phrases in the source language. In this work, we do not have this task because we use a ready human-built monolingual specialized dictionary as a source of phrases. Thus, our main interest is to develop a procedure that would provide the highest possible translation quality.

To build a phrase table, we used the Russian-English sub-corpus of CCMatrix dataset (v1) (Schwenk et al., 2021) downloaded from OPUS (Tiedemann, 2012). The size of the sub-corpus is approximately 140 million sentences. We aligned the words in the parallel corpus using GIZA++ with the 'grow-diag-final-and' heuristic. The default configuration of the Moses pipeline[1] (Koehn et al., 2007) was used to produce a phrase table.

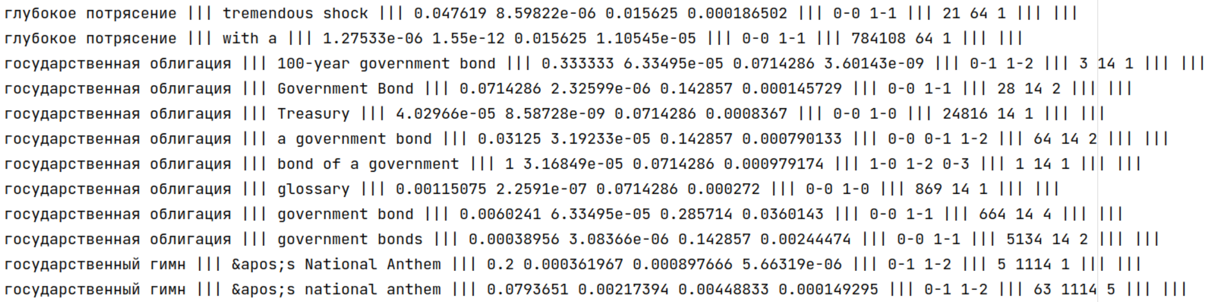
The excerpt of the resulting phrase table is given in Figure 1. For any source phrase there is a number of translation candidates along with scores, word alignments, and counts. Let us denote English phrase as e, and foreign (Russian in our case) phrase as f. Then three counts are given:

$count(e)$, number of times e was identified as a phrase in a parallel corpus;

$count(f)$, number of times f was identified as a phrase in a parallel corpus;

$count(e, f)$, number of times phrase e was translated as phrase f.

[1] https://www.statmt.org/moses/



```
глубокое потрясение ||| tremendous shock ||| 0.047619 8.59822e-06 0.015625 0.000186502 ||| 0-0 1-1 ||| 21 64 1 ||| |||
глубокое потрясение ||| with a ||| 1.27533e-06 1.55e-12 0.015625 1.10545e-05 ||| 0-0 1-1 ||| 784108 64 1 ||| |||
государственная облигация ||| 100-year government bond ||| 0.333333 6.33495e-05 0.0714286 3.60143e-09 ||| 0-1 1-2 ||| 3 14 1 ||| |||
государственная облигация ||| Government Bond ||| 0.0714286 2.32599e-06 0.142857 0.000145729 ||| 0-0 1-1 ||| 28 14 2 ||| |||
государственная облигация ||| Treasury ||| 4.02966e-05 8.58728e-09 0.0714286 0.0008367 ||| 0-0 1-0 ||| 24816 14 1 ||| |||
государственная облигация ||| a government bond ||| 0.03125 3.19233e-05 0.142857 0.000790133 ||| 0-0 0-1 1-2 ||| 64 14 2 ||| |||
государственная облигация ||| bond of a government ||| 1 3.16849e-05 0.0714286 0.000979174 ||| 1-0 1-2 0-3 ||| 1 14 1 ||| |||
государственная облигация ||| glossary ||| 0.00115075 2.2591e-07 0.0714286 0.000272 ||| 0-0 1-0 ||| 869 14 1 ||| |||
государственная облигация ||| government bond ||| 0.0060241 6.33495e-05 0.285714 0.0360143 ||| 0-0 1-1 ||| 664 14 4 ||| |||
государственная облигация ||| government bonds ||| 0.00038956 3.08366e-06 0.142857 0.00244474 ||| 0-0 1-1 ||| 5134 14 2 ||| |||
государственный гимн ||| &apos;s National Anthem ||| 0.2 0.000361967 0.000897666 5.66319e-06 ||| 0-1 1-2 ||| 5 1114 1 ||| |||
государственный гимн ||| &apos;s national anthem ||| 0.0793651 0.00217394 0.00448833 0.000149295 ||| 0-1 1-2 ||| 63 1114 5 ||| |||
```

Figure 1: The excerpt of the phrase table generated from the Russian-English sub-corpus of CCMatrix dataset.

Based on these counts the probability scores are calculated as:

$p(f|e) = count(e, f)\ /\ count(e)$, inverse phrase translation probability;

$p(e|f) = count(e, f)\ /\ count(f)$, direct phrase translation probability.

We are interested in $count(e, f)$ and probabilities $p(f|e)$, $p(e|f)$.

3.1 Selecting Translations

The process of selecting translations is as follows. We first sort all the candidates by their $count(e, f)$, which is the number of times two phrases appear to be translations of each other, and take the top 10 candidates. This is equivalent to sorting by $p(e|f)$, since $count(f)$ is the same number for a given source phrase. We then filter these candidates using thresholds. First, we filter by direct phrase translation probability $p(e|f)$, then by inverse phrase translation probability $p(f|e)$, and finally by $count(e, f)$.

We found out empirically that setting $p(e|f)$ threshold based on counts leads to better results compared to using a single universal threshold. The threshold for direct phrase translation probability $p(e|f)$ should be inversely related to $count(f)$: the more times a phrase appears in a corpus, the more appropriate translations will be identified and thus their individual probabilities will be lower. With this in mind, we set gradual thresholds for $p(e|f)$: from 0.2 for $count(f) < 50$ down to 0.04 for $count(f) > 1000$.

We also set a threshold for $p(f|e)$ to 0.04 because this helps to filter out the common type of wrong translations: when a phrase is translated as some irrelevant but highly frequent phrase or, more often, word as 'the', 'to', etc. In this case, the probability $p(e|f)$ can be very high, since the alignment error is systematic, but $p(f|e)$ is usually near $10e - 5$. We set the threshold higher than this to also get rid of translations that are not exactly wrong but rather incomplete, for example: 'inspiration' instead of 'source of inspiration'.

Additionally, we set a threshold for $count(e, f)$ to 3 since we want any phrase to occur at least 3 times with a given translation.

It might sometimes happen that none of the candidates satisfies these thresholds. In this case, we gradually lower the thresholds such that at each step there is at least one candidate remaining.

The values we select for thresholds are not optimal, but they were chosen based on the analysis of scores and counts of translations for randomly sampled phrases with different counts.

3.2 Post-processing

Finally, when we have a list of translation candidates, we clean it by removing near duplicates. First, we lower-case all candidates. We did not lower-case the corpora before feeding it to GIZA++, so there might be same translations but in different casing, e.g., 'Stock Exchange' and 'stock exchange'. Second, we detokenize the candidates because the output is still Moses-tokenized. Third, we strip (trim) punctuation from both sides, because very often we can get options like: ‘in a sense**,**’ and ‘**,** in a sense**,**’. With lowercasing and stripped punctuation, we can already get rid of some duplicates. The next step is to group same translations which come with different articles (‘a’, ‘an’, ‘the’) and phrases with infinitives that may start with or without ‘to’ preposition, e.g.: ‘**to** pave the way’ and ‘pave the way’. After grouping, we choose the one preferred form and remove the others.

As a result, we obtain a refined list of sorted translations—one-two on average for every source phrase.

4 Data

We took the manually constructed dictionary[2] of n-gram lexical units from Russian National Corpus as a source of phrases for our bilingual dictionary. Namely, it is a compilation of Russian stable lexical phrases grouped by the functions they perform:

(1) prepositions (190), e.g.:
согласно с (soglasno s) ‘in accordance with’,
во имя (vo imja) ‘in the name of’;

(2) adverbs and predicatives (2164), e.g.:
в итоге (v itoge) ‘ultimately’,
в двух словах (v dvuh slovah) ‘in a nutshell’;

(3) conjunctions and connective words (59), e.g.:
а именно (a imenno) ‘namely’,
если бы (esli by) ‘if only’;

(4) particles (24), e.g.:
едва не (edva ne) ‘nearly’,
как раз (kak raz) ‘exactly’;

(5) comment clauses (194), e.g.:
без сомнения (bez somnenija) ‘undoubtedly’,
грубо говоря (grubo govorja) ‘roughly speaking’.

We manually removed some phrases from the original dictionary, e.g., the ones which are non-contiguous or too rare. The final number of phrases in each group is indicated in brackets.

We also introduce one more **golden truth dictionary** of Russian-English phrases we built manually to evaluate our approach. We took the first 30 pages of the online Russian-English collocations dictionary[3] as a basis and updated, removed, and added some translations. Mainly, we were replacing some uncommon translations with more common ones and unifying phrase forms. The resulting dictionary consists of various phrase types, including noun phrases (‘double agent’), phrasal verbs (‘tear apart’), idiomatic expressions (‘guinea pig’), comment clauses (‘to put it mildly’), etc. Overall, there are 250 entries in the dictionary.

5 Results and Analysis

We first evaluate our approach to translating phrases using the golden truth dictionary that we built. Using our methodology, we obtain translations for each source (Russian) phrase in the dictionary if it is found in the phrase table. Out of 250 phrases, 241 were found and 9 were missing. We consider missing phrases as wrong when calculating the overall translation accuracy. We use two evaluation modes: *top1* mode, where only the first (best) translation is assessed, and *any* mode, where a phrase is considered as translated correctly if at least one of its translations matches the reference.

To have a baseline, we translated the same dictionary with a pretrained Russian-English MarianMT neural translation model (Tiedemann and Thottingal, 2020) implemented in Transformers library[4]. This

[2] `https://ruscorpora.ru/new/obgrams.html`
[3] `https://audio-class.ru/english-collocations/vocabulary-02.php`
[4] `https://huggingface.co/docs/transformers/model_doc/marian`

Method	**Accuracy (%)**
Our, *any*	69.2
Our, *top1*	62.4
NMT	38.4

Table 1: Accuracy of phrase translations measured against the golden truth dictionary. *Our* is our phrase table based method and *NMT* is a baseline method where translations are obtained from MarianMT model.

$count(f)$	**# phrases**	**Accuracy (%)**
< 10	12	25.1
10 - 50	26	69.2
50 - 100	15	86.6
100 - 200	24	62.5
200 - 500	29	82.7
500 - 1k	39	79.1
1k - 5k	50	80.2
5k - 50k	32	56.6
> 50k	14	78.3

Table 2: Accuracy of phrase translations measured against the golden truth dictionary depending on source phrase counts, $count(f)$.

model (opus-mt-ru-en[5]) was trained on combined Russian-English datasets from OPUS, where CCMatrix is a major one. The same way as with phrase table candidates, we stripped the punctuation from translations. Here, there is always just one translation for any phrase.

We lower-cased both candidate and reference translations and considered a translation correct if it matches the reference as-is or after being adjusted for articles and prepositions ('a', 'the', 'an', 'to'). To clarify, we regard 'a stray dog'/' the stray dog'/'stray dog' or 'to commit a crime'/'commit a crime' as equivalent translations.

The evaluation results are presented in Table 1. We see that regardless of the mode (*top1*/*any*), translations obtained using phrase table are significantly more accurate than the ones we got plainly translating using MarianMT, and the difference is at least 24%. We suppose the main reason for the low NMT performance is that the model is not trained to translate phrases, instead being trained on full sentences.

If we take a closer look at the results (Table 3), we will see that in the majority of cases we get correct translations (rows 1-4) for different phrase types: noun phrases ('tough stance'), idioms ('scapegoat'), comment clauses ('simply put'), etc. Sometimes there is more than one candidate, and mostly they represent valid alternatives, e.g., 'simply put' and 'in simple terms'.

The next four rows (5-8) in Table 3 showcase translation candidates that are valid although they do not match the reference. The phrases 'at a loss', 'in disbelief' are synonymous with the word 'puzzled' (row 6); and 'in the first place' (row 5) is actually even more accurate translation for the source phrase than the reference is. The last row illustrates the frequent case when the translation candidate differs from the reference by added preposition or article ('**a** full set **of**').

Let us now turn to more problematic cases demonstrated in Table 4. The first row shows how the main term ('gist') is being lost during translation. This can be attributed to the low phrase count. The next example (row 2) illustrates the challenging case where the source phrase may have several meanings depending on the context. If we consider the source phrase as complete, then the correct translation will be the reference one, 'one by one'. However, if it is a part of the bigger phrase, e.g. 'по одному поводу'

[5] `https://huggingface.co/Helsinki-NLP/opus-mt-ru-en`

	Source phrase	Candidate translations	Reference translation	$count(f)$	Correct
1	в рамках бюджета v ramkah bjudzheta	within budget, on budget, within the budget, under budget	within budget	957	Yes
2	козёл отпущения kozjol otpushhenija	scapegoat	scapegoat	31	Yes
3	проще говоря proshhe govorja	simply put, to put it simply, in simple terms	simply put	9389	Yes
4	жёсткая позиция zhjostkaja pozicija	tough stance	tough stance	49	Yes
5	в первую очередь v pervuju ochered’	primarily, in the first place, first of all	first and foremost	102472	+-
6	в недоумении v nedoumenii	at a loss, in disbelief	puzzled	1108	+-
7	время от времени vremja ot vremeni	from time to time, occasionally	once in a while	36744	+-
8	полный комплект polnyj komplekt	complete set of, a full set of	full set	1473	+-

Table 3: Phrase translation examples for the test dictionary. The candidates are valid, even if they do not match the reference.

	Source phrase	Candidate translations	Reference translation	$count(f)$	Correct
1	суть рассказа sut’ rasskaza	the story	gist of the story	7	No
2	по одному po odnomu	on one	one by one	18409	No
3	устье реки ust’e reki	the mouth of the, the mouth of the river	river mouth	876	+-
4	ни с того ни с сего ni s togo ni s sego	no apparent reason	without any rhyme or reason	210	No
5	аллергия на пыльцу allergija na pyl’cu	are allergic to pollen	pollen allergy	119	+-
6	подопытный кролик podopytnyj krolik	the experimental rabbit	guinea pig	9	No

Table 4: Phrase translation examples for the test dictionary. The candidates are partially valid or wrong.

(po odnomu povodu) meaning ‘on one occasion’, then the suggested ‘on one’ translation is the correct one.

Rows 3 and 4 exemplify the problem of inappropriately trimmed translations: ‘the mouth of the’ lacks the defining word ‘river’; ‘no apparent reason’ should start with the preposition ‘for’. In row 5, ‘are allergic to pollen’ carries the correct meaning but has a wrong form, whereas ‘the experimental rabbit’ is an uncommon translation of the Russian phrase that is best translated as ‘guinea pig’. The count (9) in the latter case is quite low, though.

Phrase counts for valid translations shown in Table 3 differ from 31 to 102k, yet, there are even less frequent phrases translated correctly, for example, ‘turnkey business’ with only 9 occurrences. However, if a phrase is very rare, the chances to get a good translation are low. We measured accuracy for phrases with different source counts in Table 2. We see the drastic decrease in accuracy for phrases with $count(f) < 10$, which suggests that 10 can be used as a default threshold when automatically constructing a dictionary. It is also interesting to note that the increase in count does not necessarily imply the increase in accuracy.

To sum up, we see many good translations, sometimes with a fair choice of options. Even if translations do not match the reference, they are mostly valid alternatives. Sometimes the translations are strangely trimmed and have an improper form or represent an uncommon translation. With all that, we almost do not observe any completely irrelevant translations after the performed filtering and post-processing.

Turning to the NMT phrase translations, we see a number of problems. One of them is word-by-word translations of idiomatic expressions: ‘single wolf’ instead of ‘lone wolf’, ‘beating of infants’ instead of ‘massacre of the innocents’, ‘aerial snakes’ for ‘kite’, etc. There are also many sub-optimal translations like ‘eastern kitchen’ for ‘oriental cuisine’ and ‘artistic literature’ for ‘fiction’ due to the literal translation of the phrases. The other problem is unexpected, lengthy translations: ‘i don’t know what i’m talking about’ for ‘pick the nose’, ‘well, let’s just put it that way’ for ‘simply put’, and so forth. Most likely, this happens because the model is trained to produce full sentences. One more important limitation is that the model cannot produce alternative options. Even if beam search with several outputs is used, the variation in translations is quite low.

Let us now focus on the generated bilingual dictionary and assess its overall practical utility. We first note that we set a threshold on $count(f)$, the number of times a source phrase appeared in a corpus, following the above analysis. We set this threshold to minimum 10 occurrences. As a result, from 1% to 26% of phrases, depending on the group, were excluded from the final dictionary.

We went through the resulting translations, and we can say that we are mostly satisfied with the resulting quality. The most common problems we noticed are the ones connected to the phrase context, as with ‘one by one’ example above. Specifically, some phrases, especially if they are short, should have different translations if they are considered a part of a bigger phrase and if they are considered a complete phrase on their own.

Apart from that, we see that very often good alternatives do not survive filtering by thresholds. Obviously, there is a trade-off between recall and precision, and we choose the latter. A potential solution that can lead to the best possible quality is to use this dictionary (and our method in general) as a basis for manual dictionary creation. Such approach saves a tremendous amount of time and effort required for the search of appropriate translations. Even if individual candidates are a bit noisy and strangely trimmed (like ‘good as it gets’), they can come as a tip for a dictionary creator pointing to the right translation (‘**as** good as it gets’). This work can be performed by language enthusiasts in a crowd-sourcing manner, for example. In this case, the thresholds should be lowered further such that rare but correct translations do not get omitted. We would like to note that looking at full lists of candidates in a phrase table is not realistic—often there are hundreds of quite irrelevant options.

Another possible improvement is extending the dictionary with parallel sentence examples showcasing a given translation option (highlighting the aligned phrases in a source and target sentences). This can be implemented if the parallel corpus used for phrase table creation is available.

Overall, we evaluate the resulting bilingual Russian-English specialized phrase dictionary as a useful resource for those whose first language is Russian and who learn/use English as a second language. Especially, it can be helpful for those who write in English and has a frequent need to translate common introductory, connective, adverbial, and other above-mentioned types of phrases from Russian to English. We note, however, that this dictionary should be used only as a source of translation options, which should be checked elsewhere if a person is unsure, keeping in mind the automatic nature of this language resource. It also can be used by those who work on creating language learning tools and writing assistants as a raw resource for further processing.

As for the approach in general, we believe that despite its simplicity, it is one of the most affordable ways to automatically compile a bilingual phrase dictionary of decent quality. It can be particularly useful in the low-resource setting, where manually created resources do not exist or are incomplete but there is a parallel corpus available. The minimal size requirements for such corpus is, however, an open research question.

To make our study complete, we need to note that although we did not need it in this work, the important aspect of the automatic dictionary creation is the automatic extraction of meaningful phrases/collocations from text corpora. There exist a number of approaches for this task (Pecina, 2005; Bhalla and Klimcikova, 2019) and we think that their choice depends on the type and the purpose of the dictionary one wants to create.

The constructed dictionary in its current form is publicly available at `https://github.com/bilingual-phrase-dict/ru-en`.

6 Conclusion

This work raises an important issue of phrase translation. We emphasize the need for high-quality phrase translation models for second language learners and users and suggest a simple approach for obtaining phrase translations based on the GIZA++ output. Using this approach, we automatically construct a new Russian-English bilingual phrase dictionary and make it publicly available. We analyze the quality of our approach and highlight its strengths and shortcomings. We also compare it to translating phrases with a state-of-the-art neural machine translation model and show how poor NMT model performs in translating phrases. We see this as a problem and expect that future research will address it by proposing high-quality phrase translation models.

References

Vishal Bhalla and Klara Klimcikova. 2019. Evaluation of automatic collocation extraction methods for language learning. // *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, P 264–274, Florence, Italy, August. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311, jun.

Yi-Jyun Chen, Ching-Yu Helen Yang, and Jason S. Chang. 2020a. Improving phrase translation based on sentence alignment of Chinese-English parallel corpus. // *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, P 6–7, Taipei, Taiwan, September. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020b. Accurate word alignment induction from neural machine translation. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 566–576, Online, November. Association for Computational Linguistics.

Andrejs Veisbergs Daiga Deksne. 2018. A workflow for supplementing a latvian-english dictionary with data from parallel corpora and a reversed english-latvian dictionary. // *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, P 127–135, Ljubljana, Slovenia, jul. Ljubljana University Press, Faculty of Arts.

Marcos Garcia, Marcos García-Salido, and Margarita Alonso-Ramos. 2019. Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics.

Zhang Jun. 2008. A comprehensive review of studies on second language writing. *HKBU Papers in Applied Language Studies*, 12(2).

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. *International Workshop on Spoken Language Translation*, 01.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. // *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, P 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. // *31st Annual Meeting of the Association for Computational Linguistics*, P 17–22, Columbus, Ohio, USA, June. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 03.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. // *Proceedings of the ACL Student Research Workshop*, P 13–18, Ann Arbor, Michigan, June. Association for Computational Linguistics.

John Richardson, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Bilingual dictionary construction with transliteration filtering. // *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, P 1013–1017, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Oscar Mendoza Rivera, Ruslan Mitkov, and Gloria Corpas Pastor. 2013. A flexible framework for collocation retrieval and translation from parallel and comparable corpora. // *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technologies*, Nice, France, September 3.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, P 6490–6500, Online, August. Association for Computational Linguistics.

Violeta Seretan and Éric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. // *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, P 375–384, Toulouse, France, June. ATALA.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Comput. Linguistics*, 22:1–38.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. // *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, P 479–480, Lisboa, Portugal, November. European Association for Machine Translation.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. // *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, P 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Zorana Vasiljevic. 2014. Teaching collocations in a second language: Why, what and how. *Elta Journal*, 2(2):48–73.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 1605–1617, Online, July. Association for Computational Linguistics.