УДК 004.912 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2022/4/189-201
Поступила в редакцию 12.04.2022
Подписана в печать 05.12.2022

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ДВУЯЗЫЧНОГО СЛОВАРЯ НА ОСНОВЕ ВЫВОДА GIZA++

© 2022 А. М. Хусаинова✉, В. А. Романов, А. М. Хан

Университет Иннополис
ул. Университетская, 1, 420500 Иннополис, Российская Федерация

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 статистического инструмента GIZA++ с последующей фильтрацией с помощью эвристик. Мы анализируем качество перевода, полученного с помощью данного подхода, и сравниваем его с эталонным переводом и с переводом фраз с помощью системы НМП, обученной на предложениях. Результаты показывают, что несмотря на выявленные проблемы, переводы фраз чаще всего корректны, и даже если они не совпадают с эталонным переводом, представляют собой возможные альтернативные переводы. Также важным результатом является то, что данный подход работает значительно лучше, чем перевод фраз с помощью системы НМП. Используя предложенный подход, мы получили русско-английский словарь лексических оборотов, который можно использовать как в готовом виде, так и в качестве исходного материала для составления словаря вручную. Полученный русско-английский фразовый словарь был размещен в сети Интернет в качестве лингвистического ресурса.
Ключевые слова: перевод фраз, перевод коллокаций, машинный перевод, автоматическое построение словаря, двуязычный словарь, фразовый словарь, языковые ресурсы.

ВВЕДЕНИЕ

Изучающие иностранный (второй) язык обычно используют свой родной (первый) язык для поиска перевода слов, фраз и предложений на втором языке. Иногда необходимо перевести готовые предложения, например, когда пользователь сначала составляет их на родном языке. Однако, когда пользователь сразу формирует предложение на втором языке, ему часто приходится обращаться к словарю для правильного перевода слова или фразы, что особенно актуально при написании текста на втором языке [1].

Формирование словарного запаса — это базовый этап в изучении любого языка. Однако, знания отдельных слов недостаточно. Чаще всего именно фразы, а не слова, играют роль смысловых единиц, поэтому так важно изучение словосочетаний [2]. Поэтому хорошие инструменты для изучения языка всегда предлагают для изучения пользователю слова и фразы вкупе, чтобы он мог их понимать и формировать на их основе связные предложения. Таким образом, для создания инструментов для изучения языка необходимо иметь не только обычные словари, где базовая единица — слово, но и качественные фразовые словари.

С другой стороны, у тех, кто уже владеет и пользуется иностранным языком, также во

✉ Хусаинова Альбина Маратовна
e-mail: a.khusainova@innopolis.ru

многих ситуациях возникает необходимость в переводе фраз. Например, при чтении текстов, в которых встречаются незнакомые слова или фразы. Хороший пример — это электронные книги со всплывающими подсказками при наведении на слово или выражение. Пользователей может интересовать перевод непосредственно фразы или, если речь идет о незнакомом слове, они могут захотеть узнать распространенные словосочетания с данным словом вместе с их переводами, что также ведет к словарям фраз.

Другой пример — это написание текста на иностранном языке. Когда пользователь формирует предложение, он либо сразу вспоминает нужные слова и словосочетания, либо ему приходится переводить их с родного языка. В последнем случае очень полезно предложить пользователю набор возможных переводов, чтобы он мог выбрать тот, который несет в себе задуманный смысл и лучше всего соответствует контексту. Предоставление таких вариантов возможно только при наличии соответствующих языковых ресурсов.

Переводы слов как правило можно найти в словарях, составленных вручную, а переводы предложений обычно можно получить с помощью онлайн-инструментов НМП. Однако, когда речь идет о фразах, ситуация обстоит сложнее. Чаще всего хорошие двуязычные словари частоупотребимых выражений имеются только у высокоресурсных языковых пар. Даже если они есть, такие словари зачастую неполные или ограниченные, например, именными словосочетаниями. Что касается нейронного перевода, то модели, обученные на целых предложениях, часто не дают качественного результата для фраз — перевод может быть просто ошибочным или это может быть единственный перевод, в то время как на самом деле существует множество одинаково хороших альтернатив. Иногда эта проблема решается путем включения информации из существующих словарей — когда пользователь ищет перевод распространенного выражения, система переключается с нейронного перевода на простой поиск в словаре. Однако, как уже упоминалось, для многих языковых пар таких словарей в принципе не существует.

В данной работе мы предлагаем подход для автоматического создания двуязычного словаря фраз на основе корпуса параллельных предложений. Мы получаем переводы-кандидаты из таблицы фраз, которая является результатом работы статистического инструмента GIZA++ [3, 4], а затем фильтруем и сортируем их с помощью эвристик. В результате мы получаем словарь фраз, который может быть использован в исходном виде либо в качестве основы для составления словаря вручную. Мы исследовали получившийся словарь и оценили его качество в сравнении с эталонным и нейронным переводом. Наконец, мы разместили созданный русско-английский фразовый словарь в сети Интернет в качестве лингвистического ресурса.

Анализ предшествующих работ

Задача перевода фраз как таковая не представлена в литературе. Однако есть некоторые, в основном старые, работы по переводу *коллокаций*. Поскольку термин *коллокация* очень близок по смыслу к термину фраза, мы считаем литературу по переводу коллокаций релевантной. В наиболее недавней работе [5] предлагается использовать вектора слов для поиска двуязычных коллокаций — сначала сопоставляются ключевые слова коллокаций, а затем варианты их зависимых слов. Недостатком такого подхода является то, что он ограничивает перевод коллокаций только очень точными соответствиями, в то время как довольно часто фразы могут быть более идиоматичными. Кроме того, согласно данному подходу, количество слов в коллокации должно соответствовать количеству слов в ее переводе, что также часто не соответствует действительности. Например, английская фраза “bring about” может быть переведена на русский язык одним словом «вызывать».

Что касается предыдущих работ, в [6] переводят коллокации пословно максимизируя значения коэффициента Дайса между исходными и целевыми коллокациями в параллельном корпусе. Авторы делают не очень реалистичное предположение, что любая коллокация имеет уникальный перевод. Ана-

логичным образом, в [7] сначала выделяют именные словосочетания на двух языках, а затем максимизируют их совместную встречаемость используя двуязычный корпус.

В работе [8] предполагается, что коллокации в обоих языках имеют одинаковую структуру по частям речи. Используя словари, они находят перевод для ключевого слова, а затем ищут соответствующие коллокации с той же структурой в предложениях параллельного корпуса. В [9] используется аналогичный подход — ключевые слова переводятся с помощью словаря, а для поиска соответствующих коллокаций применяется синтаксический разбор.

В нашем случае не предполагается, что фразы имеют одинаковую синтаксическую или частеречную структуру. Кроме того, поскольку мы рассматриваем не только коллокации, выбор ключевого слова может быть неоднозначным. Следовательно, мы не рассматриваем подходы, которые основаны на сопоставлении ключевых слов и полагаются синтаксическое / частеречное соответствие.

Вместо этого мы склоняемся к методам, которые находят переводы фраз с помощью выравнивания слов в параллельных текстах. Одним из самых эффективных статистических инструментов для выравнивания слов в параллельных текстах является GIZA++ [3,4]. Несмотря на то, что лежащие в ее основе технологии были разработаны десятилетия назад, современные нейронные методы до сих пор не могут полностью превзойти GIZA++. Лишь недавно некоторые работы [10,11], использующие нейронные архитектуры, смогли показать некоторые улучшения по сравнению с GIZA++. Однако анализ показывает, что эти улучшения связаны с повышением полноты, но не точности. В нашем случае точность важнее, так как при составлении словаря лучше иметь меньшее количество более точных результатов.

После выравнивания слов в обоих направлениях перевода полученные результаты объединяются с помощью метода “grow-diag” [12]. Далее фразы извлекаются на основе критерия консистентности: «Слова в паре фраз должны быть сопоставлены друг с другом (согласно выравниванию) и не должны быть сопоставлены с другими словами предложения» [12]. В результате получается список фраз с их возможными переводами и вероятностями этих переводов. Такой список называется таблицей фраз и изначально задумывался как часть системы статистического машинного перевода. В настоящее время статистический машинный перевод заменен нейронным машинным переводом, однако этот побочный продукт, таблица фраз, все еще оказывается полезным.

Работы, похожие на нашу, которые используют таблицы фраз для создания / расширения двуязычных словарей, включают [13, 14] и [15]. В следующем разделе мы подробно описываем наш подход.

1. МЕТОДЫ И МАТЕРИАЛЫ

Наша цель — построить фразовый словарь, и нам необходимо определить, что мы подразумеваем под *фразой*. Мы понимаем фразы как n-граммы, которые несут определенный смысл, встречаются вместе чаще, чем просто случайно (как коллокации), и общий смысл которых необязательно может быть понятен из отдельных слов (как идиомы). Необходимо отметить, что из-за ограничения выбранного метода выравнивания мы рассматриваем только неразрывные фразы.

Первый шаг при составлении двуязычного словаря — это определение словосочетаний на исходном языке. В данной работе такая задача перед нами не стоит, поскольку в качестве источника фраз мы используем существующий словарь лексических оборотов. Таким образом, наш основной интерес заключается в разработке метода, который обеспечит максимально возможное качество перевода.

Для построения таблицы фраз мы использовали русско-английский подкорпус набора данных CCMatrix (v1) [16], размещенный на сайте OPUS [17]. Размер подкорпуса составляет около 140 миллионов предложений. Мы выровняли слова в параллельном корпусе с помощью GIZA++ применив эвристику “grow-diag-final-and”. Для создания таблицы

фраз использовалась система Moses [18, 19] в конфигурации по умолчанию.

Фрагмент полученной таблицы фраз приведен на рис. 1. Для каждой фразы на исходном языке обычно существует множество вариантов перевода, для каждого из которых представлены вероятности, соответствия слов согласно выравниванию и показатели встречаемости. Обозначим английскую фразу как e, а русскую фразу как f. Тогда мы имеет три показателя встречаемости:

$count(e)$, количество раз, когда e была определена как фраза в параллельном корпусе;

$count(f)$, количество раз, когда f была определена как фраза в параллельном корпусе;

$count(e, f)$, количество раз, когда фраза e была переведена как фраза f.

На основе этих показателей рассчитываются вероятности: $p(f \mid e) = count(e, f) / count(e)$, обратная вероятность перевода фразы;

$p(e \mid f) = count(e, f) / count(f)$, прямая вероятность перевода фразы.

Нам понадобятся $count(e, f)$ и вероятности $p(f \mid e)$, $p(e \mid f)$.

1.1. Выбор вариантов перевода

Процесс выбора вариантов перевода происходит следующим образом. Сначала мы сортируем кандидатов по $count(e, f)$, то есть по количеству раз, когда две фразы определялись как переводы друг друга, и отбираем 10 лучших кандидатов. Это эквивалентно сортировке по $p(e \mid f)$, поскольку $count(f)$ одинаково для всех переводов данной фразы f. Затем мы фильтруем полученный список кандидатов с помощью пороговых значений. Сначала мы фильтруем по прямой вероятности перевода $p(e \mid f)$, затем по обратной вероятности $p(f \mid e)$ и, наконец, по $count(e, f)$.

Эмпирическим путем мы выяснили, что задание порога $p(e \mid f)$ на основе показателей встречаемости приводит к лучшим результатам по сравнению с использованием единого фиксированного порога. Порог прямой вероятности перевода $p(e \mid f)$ должен быть в обратной зависимости от $count(f)$: чем чаще фраза встречается в корпусе, тем больше будет выявлено подходящих переводов и, следовательно, их индивидуальные вероятности будут ниже. Исходя из этого, мы установили ступенчатые пороги для $p(e \mid f)$: от 0.2 для $count(f) < 50$ до 0.04 для $count(f) > 1000$.

Мы также установили порог для $p(f \mid e)$ равным 0.04, поскольку это помогает отсеять распространенный тип неправильных переводов, когда фраза переводится как нерелевантная, но очень часто встречающаяся фраза или чаще слово, такое как “the”, “to” и т. д. В этом случае вероятность $p(e \mid f)$ может быть очень высокой, поскольку ошибка выравнивания часто бывает систематической, но $p(f \mid e)$ обычно составляет порядка 10e–5. Мы установили порог выше этого значения, чтобы также избавиться от переводов, которые не совсем неправильные, но скорее неполные, например, “inspiration” вместо “source of inspiration”.

Кроме того, мы установили порог для $count(e, f)$ равным 3, чтобы фраза с определенным переводом встречалась как минимум 3 раза.

Иногда может случиться так, что ни один из вариантов перевода не удовлетворяет этим

```
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 ||| |||
```

Рис. 1. Фрагмент таблицы фраз, полученной на основе русско-английского подкорпуса набора данных CCMatrix

[Fig. 1. The excerpt of the phrase table generated from the Russian-English sub-corpus of CCMatrix]

пороговым значениям. В таком случае мы постепенно снижаем пороговые значения так, чтобы на каждом шаге оставался хотя бы один кандидат.

Значения пороговых значений не являются оптимальными, однако, они были определены на основе анализа вероятностей и показателей встречаемости переводов случайно отобранных фраз разной частоты.

1.2. Постобработка

Теперь, имея сокращенный список кандидатов перевода фразы, мы очищаем его, удаляя около-дубликаты. Во-первых, мы переводим все кандидаты в нижний регистр. Мы не переводили корпус в нижний регистр перед тем, как подавать его в GIZA++, поэтому возможны одинаковые переводы в разном регистре, например, “Stock Exchange” и “stock exchange”. Во-вторых, мы производим детокенизацию кандидатов, так как они все еще токенизированы Moses. В-третьих, мы убираем пунктуацию с обеих сторон фразы, так как очень часто встречаются кандидаты типа: “in a sense**,**” и “**,** in a sense**,**”. Приведение к одному регистру и удаление пунктуации уже отсекает некоторые дубликаты. Следующий шаг — это объединение одинаковых переводов с различными артиклями (“a”, “an”, “the”), а также фраз с инфинитивами, которые могут начинаться с предлогом “to” или без него, например: “**to** pave the way” и “pave the way”. После определения таких переводов мы выбираем одну предпочтительную форму и удаляем остальные.

В результате мы получаем обработанный и отсортированный список переводов — в среднем по одному-два перевода для каждой фразы.

1.3. Данные

В качестве источника фраз для нашего двуязычного словаря мы взяли существующий составленный вручную словарь [20] устойчивых лексических оборотов из Национального корпуса русского языка. Лексические обороты в этом словаре организованы по группам согласно выполняемым ими функциям:

1. *Обороты в функции предлога* (190), например: “согласно с”, “во имя”;

2. *Наречные и предикативные обороты* (2164), например:“в итоге”, “в двух словах”;

3. *Обороты в функции союза и союзного слова* (59), например: “а именно”, “если бы”;

4. *Обороты в функции частиц*, например: “едва не”, “как раз”;

5. *Вводные обороты* (194), например: “без сомнения”, “грубо говоря”.

Мы вручную удалили некоторые фразы из исходного словаря, например, редкие и не непрерывные фразы. В скобках указано окончательное количество фраз в каждой группе.

Мы также вводим еще один **эталонный словарь** русско-английских фраз, который мы составили вручную, чтобы оценить качество нашего подхода. Мы взяли за основу первые 30 страниц онлайн-словаря [21] русско-английских словосочетаний и обновили, удалили и добавили некоторые переводы. В основном, мы заменяли некоторые нетипичные переводы на более распространенные и приводили фразы к единому виду. Получившийся словарь состоит из различных типов фраз, включая именные словосочетания (“двойной агент”), идиоматические выражения (“подопытный кролик”), вводные конструкции (“мягко говоря”) и т. д. Всего в словаре 250 фраз.

2. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Сначала мы оцениваем предложенный подход к переводу фраз с помощью составленного нами эталонного словаря. Применив наш подход, мы получили перевод для каждой исходной (русской) фразы в словаре, кроме тех, которые отсутствовали в таблице фраз (9 из 250). При расчете общей точности перевода мы считали отсутствующие фразы как ошибочные. Использовалось два режима оценки: *top1*, где оценивается только первый (лучший) перевод, и *any*, где фраза считается переведенной правильно, если хотя бы один из ее переводов совпадает с эталонным.

Для сравнения мы перевели тот же словарь с помощью предобученной русско-английской нейронной модели перевода MarianMT [22], реализованной в библиотеке Transformers [23]. Эта модель (opus-mt-ru-en [24]) была обучена на комбинированных русско-английских данных из OPUS, где CCMatrix — это основной источник. Так же, как и в случае с кандидатами из таблицы фраз, мы убрали лишнюю пунктуацию из переводов. В данном случае для любой фразы всегда существует только один перевод.

Мы перевели в нижний регистр как полученные, так и эталонные переводы и считали перевод правильным, если он совпадал с эталонным как есть или после корректировки с учетом артиклей и предлогов (“a”, “the”, “an”, “to”). К примеру, мы считаем эквивалентными переводы “**a** stray dog” / “**the** stray dog” / “stray dog” или “**to** commit a crime” / “commit a crime”.

Результаты оценки представлены в табл. 1. Мы видим, что независимо от режима оценки (*top1 / any*), переводы, полученные с помощью таблицы фраз, значительно точнее, чем те, которые мы получили при переводе с помощью MarianMT, и разница составляет не менее 24 %. Мы полагаем, что основная причина низких результатов нейронной модели заключается в том, что она не обучена переводу фраз, так как обучается на полных предложениях.

Если мы обратимся к результатам на табл. 2, то увидим, что в большинстве случаев мы получаем правильные переводы (строки 1–4) для различных типов фраз: именные словосочетания (“tough stance”), идиомы (“scapegoat”), вводные фразы (“simply put”) и т. д. Иногда имеется несколько вариантов, и в основном они представляют собой допустимые альтернативы, например, “simply put” и “in simple terms”.

Следующие четыре строки (5–8) в табл. 2 демонстрируют случаи, когда варианты перевода, хотя и не соответствуют эталону, являются корректными. Фразы “at a loss”, “in disbelief” синонимичны слову “puzzled” (строка 6); а фраза “in the first place” (строка 5) на самом деле является даже более точным переводом для исходной фразы, чем эталонный. Последняя строка иллюстрирует частый случай, когда перевод отличается от эталона добавлением предлога или артикля (“**a** full set **of**”).

Теперь обратимся к более проблематичным случаям, продемонстрированным в табл. 3. В первая строке видно, как основной термин (“gist”) теряется при переводе. Это может быть обусловлено редкой встречаемостью фразы (7). Следующий пример (строка 2) иллюстрирует сложный случай, когда исходная фраза может иметь несколько значений в зависимости от контекста. Если рассматривать исходную фразу как полную, то правильным переводом будет эталонный — “one by one”. Однако если это часть более длинной фразы, например, “по одному поводу”, то предложенный перевод “on one” является верным.

Строки 3 и 4 иллюстрируют проблему обрезанных переводов: в переводе “the mouth of the” не хватает определяющего слова “river”; “no apparent reason” должно начинаться с предлога “for”. В строке 5 перевод “are allergic to pollen” имеет правильное значение, но неправильную форму, в то время как “the experimental rabbit” является нетипичным переводом русской фразы, которую лучше перевести как “guinea pig”. Заметим, что последняя фраза встречалась достаточно редко (9).

Показатели встречаемости фраз, которые были переведены правильно (табл. 2), варьируются от 31 до 102 тыс., однако есть и менее частотные правильно переведенные фразы, например, “бизнес под ключ”, у которого всего 9 вхождений. Все же в общем случае, если фраза очень редкая, шансы получить хоро-

*Таблица 1. Точность перевода фраз, измеренная на основе эталонного словаря. **Наш** — метод, основанный на таблице фраз, **НМП** — базовый метод, где переводы получены с помощью модели MarianMT*
[Table 2. 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]

Метод	**Точность (%)**
Наш, *any*	69.2
Наш, *top1*	62.4
НМП	38.4

Таблица 2. Примеры перевода фраз тестового словаря. Варианты перевода являются правильными, даже если они не всегда соответствуют эталонному переводу
[Table 2. Phrase translation examples for the test dictionary. The candidates are valid, even if they do not match the reference]

	Исходная фраза	**Варианты перевода**	**Эталонный перевод**	***count(f)***	**Верно**
1	в рамках бюджета	within budget, on budget, within the budget, under budget	within budget	957	Да
2	козёл отпущения	scapegoat	scapegoat	31	Да
3	проще говоря	simply put, to put it simply, in simple terms	simply put	9389	Да
4	жёсткая позиция	tough stance	tough stance	49	Да
5	в первую очередь	primarily, in the first place, first of all	first and foremost	102472	+–
6	в недоумении	at a loss, in disbelief	puzzled	1108	+–
7	время от времени	from time to time, occasionally	once in a while	36744	+–
8	полный комплект	complete set of, a full set of	full set	1473	+–

Таблица 3. Примеры перевода фраз тестового словаря. Варианты перевода частично верны либо ошибочны
[Table 3. Phrase translation examples for the test dictionary. The candidates are partially valid / wrong]

	Исходная фраза	**Варианты перевода**	**Эталонный перевод**	***count(f)***	**Верно**
1	суть рассказа	the story	gist of the story	7	Нет
2	по одному	on one	one by one	18409	Нет
3	устье реки	the mouth of the, the mouth of the river	river mouth	876	+-
4	ни с того ни с сего	no apparent reason	without any rhyme or reason	210	Нет
5	аллергия на пыльцу	are allergic to pollen	pollen allergy	119	+-
6	подопытный кролик	the experimental rabbit	guinea pig	9	Нет

ший перевод невелики. Мы измерили точность перевода фраз с разной частотностью в табл. 4. Мы наблюдаем резкое снижение точности для фраз с $count(f) < 10$, что говорит о том, что значение 10 можно использовать в качестве порога по умолчанию при автоматическом построении словаря. Также интересно отметить, что увеличение показателя встречаемости фразы не обязательно ведет к увеличению точности перевода.

В общем и целом, мы наблюдаем много хороших переводов, иногда с выбором вариантов. Даже если переводы не совпадают с эталонным, в большинстве случаев они являются корректными альтернативами. Иногда переводы странно обрезаны, имеют неподходящую форму или представляют собой нетипичный перевод. При всем этом после проведенной фильтрации и постобработки мы практически не наблюдаем совершенно нерелевантных переводов.

Перейдем к переводам фраз с помощью НМП. Здесь мы видим ряд проблем. Одна из них — дословный перевод идиоматических выражений: “single wolf” вместо “lone wolf”, “beating of infants” вместо “massacre of the innocents”, “aerial snakes” вместо “kite” и т. д. Также встречается множество неоптималь-

Таблица 4. Точность перевода фраз, измеренная на основе эталонного словаря в зависимости от показателя встречаемости исходных фраз, count(f)
[Table 4. Accuracy of phrase translations measured against the golden truth dictionary depending on source phrase counts, count(f)]

count(f)	Количество фраз	Точность (%)
< 10	12	25.1
10–50	26	69.2
50–100	15	86.6
100–200	24	62.5
200–500	29	82.7
500–1 тыс.	39	79.1
1 тыс.–5 тыс.	50	80.2
5 тыс.–50 тыс.	32	56.6
> 50 тыс.	14	78.3

ных переводов, таких как “eastern kitchen” вместо “oriental cuisine” или “artistic literature” вместо “fiction”, что происходит из-за буквального перевода фраз. Другая проблема — неожиданные, длинные переводы: “i don’t know what i’m talking about” вместо “pick the nose”, “`well, let’s just put it that way” вместо “simply put” и так далее. Скорее всего, это происходит потому, что модель обучена выдавать полные предложения. Еще одно важное ограничение заключается в том, что модель не может выдавать альтернативные варианты. Даже если использовать лучевой поиск (beam search) со множественным выводом, вариативность переводов довольно низка.

Теперь обратимся к составленному двуязычному словарю и оценим его общую практическую полезность. Прежде всего, отметим, что исходя из приведенного выше анализа, мы установили порог (10) для $count(f)$ — количества определений исходной фразы в корпусе. В результате от 1 % до 26 % фраз, в зависимости от группы, были исключены из окончательного словаря.

Мы просмотрели получившиеся переводы и можем сказать, что в целом они достаточно хорошего качества. Наиболее распространенные проблемы, которые мы заметили, связаны с контекстом фразы, как в примере с “one by one” выше. В частности, некоторые, особенно короткие, фразы должны иметь разный перевод в зависимости от того, считаются ли они частью более длинной фразы или нет.

Кроме того, мы видим, что очень часто хорошие альтернативные переводы не проходят фильтрации по пороговым значениям. Очевидно, это вопрос компромисса между полнотой и точностью, и мы выбираем последнее. Потенциальным решением, которое может привести к наилучшему качеству, является использование этого словаря (и нашего метода в целом) в качестве основы для создания словаря вручную. Такой подход позволяет сэкономить колоссальное количество времени и усилий, необходимых для поиска подходящих переводов. Даже если отдельные кандидаты немного шумные и странно обрезанные (например, “good as it gets”), они могут стать подсказкой для создателя словаря, указывающей на правильный перевод (“**as** good as it gets”). Эта работа может выполняться, например, энтузиастами в формате краудсорсинга. В этом случае пороговые значения должны быть еще ниже, чтобы более редкие, но правильные переводы не были отсеяны. Нужно отметить, что просматривать полные списки кандидатов в таблице фраз нереалистично — часто встречаются сотни совершенно нерелевантных вариантов.

Еще одно возможное улучшение — расширение словаря примерами параллельных предложений, демонстрирующих тот или иной вариант перевода (выделяя соответствующие фразы в исходном и целевом предложениях). Это может быть реализовано при наличии параллельного корпуса, использованного для создания таблицы фраз.

В целом, мы оцениваем полученный двуязычный русско-английский фразовый словарь как полезный ресурс для тех, чей родной язык — русский и кто изучает / использует английский как второй язык. Особенно он может быть полезен тем, кто пишет на английском языке и часто сталкивается с необходимостью перевода с русского на английский распространенных вводных, соеди-

нительных, наречных и других вышеупомянутых типов фраз. Отметим, однако, что данный словарь следует использовать только как источник вариантов перевода, которые следует проверять в других местах, если человек не уверен, помня об автоматической природе данного языкового ресурса. Он также может быть полезен тем, кто работает над созданием инструментов для изучения языка и помощников в написании текстов, в качестве исходного ресурса для дальнейшей обработки.

Что касается подхода в целом, мы считаем, что, несмотря на его простоту, это один из самых доступных способов автоматического составления двуязычного словаря фраз достаточно хорошего качества. Он может быть особенно полезен в условиях ограниченных ресурсов, когда созданные вручную словари не существуют либо являются неполными, но при этом имеется параллельный корпус. Однако вопрос о минимальных требованиях к размеру такого корпуса остается открытым.

Чтобы дополнить нашу работу, необходимо отметить, что, важным аспектом автоматического создания словаря является автоматическое извлечение фраз / коллокаций из текстовых корпусов. В данной работе такая задача перед нами не стояла, однако в общем случае для ее решения существует ряд подходов [25,26], и выбор конкретного подхода зависит от типа и цели создания словаря.

Созданный словарь размещен в открытом доступе по адресу https://github.com/bilingual-phrase-dict/ru-en.

ЗАКЛЮЧЕНИЕ

Наша работа поднимает важную проблему перевода фраз. Мы подчеркиваем необходимость качественных моделей фразового перевода для изучающих и использующих второй язык и предлагаем простой подход для получения перевода фраз на основе вывода GIZA++. Используя этот подход, мы автоматически составляем новый русско-английский двуязычный словарь фраз и размещаем его в открытом доступе. Мы анализируем качество нашего подхода и выделяем его достоинства и недостатки. Мы также сравниваем его с переводом фраз с помощью современной нейронной модели машинного перевода и показываем, насколько неудовлетворительно НМП модель справляется с переводом фраз. Мы считаем это проблемой и надеемся, что будущие исследования приблизят ее решение предложив новые модели для качественного перевода фраз.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Jun Z.* A comprehensive review of studies on second language writing // HKBU Papers in Applied Language Studies. – 2008. – 12.

2. *Vasiljevic Z.* Teaching collocations in a second language: Why, what and how // Elta Journal. – 2014. – 2. – P. 48–73.

3. *Brown P. F., Pietra V. J. D., Pietra S. A. D., Mercer R. L.* The Mathematics of Statistical Machine Translation: Parameter Estimation // Comput. Linguist. – 1993 Jun. – 19. – P. 263–311.

4. *Och F. J., Ney H.* A Systematic Comparison of Various Statistical Alignment Models // Computational Linguistics. – 2003 Mar. – 29. – P. 19–51. – DOI: 10.1162/089120103321337421.

5. *Garcia M, García-Salido M, Alonso-Ramos M.* Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics. – 2019.

6. *Smadja F., McKeown K., Hatzivassiloglou V.* Translating Collocations for Bilingual Lexicons: A Statistical Approach // Comput. Linguistics. – 1996. – 22. – P. 1–38.

7. *Kupiec J.* An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora // In 31st Annual Meeting of the Association for Computational Linguistics; 1993 Jun; Columbus: Association for Computational Linguistics. – P. 17–22. – DOI: 10.3115/981574.981577.

8. *Rivera O. M., Mitkov R., Corpas Pastor G.* A flexible framework for collocation retrieval and translation from parallel and comparable corpora // In Proceedings of the Workshop on

Multi-word Units in Machine Translation and Translation Technologies; 2013 Sep; Nice.

9. *Seretan V, Wehrli É.* Collocation translation based on sentence alignment and parsing. In Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs; 2007 Jun; Toulouse: ATALA. – P. 375–384.

10. *Zenkel T., Wuebker J., DeNero J.* End-to-End Neural Word Alignment Outperforms GIZA++. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul; Online: Association for Computational Linguistics. – P. 1605–1617. – DOI: 10.18653/v1/2020.acl-main.146.

11. *Chen Y., Liu Y., Chen G., Jiang X., Liu Q.* Accurate Word Alignment Induction from Neural Machine Translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov; Online: Association for Computational Linguistics. – P. 566–576. – DOI: 10.18653/v1/2020.emnlp-main.42.

12. *Koehn P., Axelrod A., Birch A., Callison-Burch C., Osborne M., Talbot D.* Edinburgh system description for the 2005 IWSLT speech translation evaluation // International Workshop on Spoken Language Translation. – 2005 Jan.

13. *Richardson J., Nakazawa T., Kurohashi S.* Bilingual Dictionary Construction with Transliteration Filtering // In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 2014 May; Reykjavik: European Language Resources Association (ELRA). – P. 1013–1017.

14. *Daiga Deksne A. V.* A Workflow for Supplementing a Latvian-English Dictionary with Data from Parallel Corpora and a Reversed English-Latvian Dictionary // In Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts; 2018 Jul; Ljubljana: Ljubljana University Press, Faculty of Arts. – P. 127–135.

15. *Chen Y. J., Yang C. Y. H., Chang J. S.* Improving Phrase Translation Based on Sentence Alignment of Chinese-English Parallel Corpus. In Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020); 2020 Sep; Taipei: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). – P. 6–7.

16. *Schwenk H., Wenzek G., Edunov S., Grave E., Joulin A., Fan A.* CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. In 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); 2021 Aug; Online: Association for Computational Linguistics. – P. 6490–6500. – DOI: 10.18653/v1/2021.acl-long.507.

17. *Tiedemann J.* Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12); 2012 May; Istanbul: European Language Resources Association (ELRA). – P. 2214–2218.

18. *Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N. [et al.]* Moses: Open Source Toolkit for Statistical Machine Translation // In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions; 2007 Jun; Prague: Association for Computational Linguistics. – P. 177–180.

19. Moses. – Режим доступа: https://www.statmt.org/moses/. (дата обращения: 06.11.2022).

20. Корпусный словарь неоднословных лексических единиц (оборотов). – Режим доступа: https://ruscorpora.ru/page/obgrams/. (дата обращения: 06.11.2022).

21. Русско-английский словарь словосочетаний и фраз. – Режим доступа: https://audioclass.ru/english-collocations/vocabulary-02.php. (дата обращения: 06.11.2022).

22. *Tiedemann J., Thottingal S.* OPUS-MT – Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation; 2020 Nov; Lisboa: European Association for Machine Translation. – P. 479–480.

23. MarianMT. – Режим доступа: https://huggingface.co/docs/transformers/model_doc/marian. (дата обращения: 06.11.2022).

24. Helsinki-NLP/opus-mt-ru-en. – Режим доступа: https://huggingface.co/Helsinki-NLP/opus-mt-ru-en. (дата обращения: 06.11.2022).

25. *Pecina P.* An Extensive Empirical Study of Collocation Extraction Methods // In Proceedings of the ACL Student Research Workshop; 2005 Jun; Ann: Association for Computational Linguistics. – P. 13–18.

26. *Bhalla V., Klimcikova K.* Evaluation of automatic collocation extraction methods for language learning. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications; 2019 Aug; Florence: Association for Computational Linguistics. – P. 264–274. – DOI: 10.18653/v1/W19-4428.

Хусаинова Альбина Маратовна — аспирант 4-го года обучения, ассистент в лаборатории машинного обучения и представления данных Университета Иннополис.
E-mail: a.khusainova@innopolis.ru
ORCID iD: https://orcid.org/0000-0002-0636-3449

Романов Виталий Анатольевич — аспирант 4-го года обучения, ассистент в лаборатории промышленной разработки ПО Университета Иннополис.
E-mail: v.romanov@innopolis.ru
ORCID iD: https://orcid.org/0000-0003-3772-0039

Хан Адил Мехмуд — кандидат физ.-мат. наук, профессор, начальник лаборатории машинного обучения и представления данных Университета Иннополис.
E-mail: a.khan@innopolis.ru
ORCID iD: https://orcid.org/0000-0003-2220-8518

DOI: https://doi.org/10.17308/sait/1995-5499/2022/4/189-201 ISSN 1995-5499
Received 12.04.2022
Accepted 05.12.2022

AUTOMATIC BILINGUAL PHRASE DICTIONARY CONSTRUCTION FROM GIZA++ OUTPUT

© 2022 A. M. Khusainova✉, V. A. Romanov, A. M. Khan

Innopolis University
1, Universitetskaya Street, 420500 Innopolis, Russian Federation

Annotation. 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. In this paper, we propose an automatic approach to create such a dictionary based on the output of the statistical tool GIZA++ followed by filtering with heuristics. We analyze the translation quality obtained with this approach and compare it with reference translations and with phrases translation using a sentences-trained NMT system. The results show that, despite the problems identified, the phrase translations are most often correct, and even if they do not match the reference translation, they represent valid alternative translations. Another important result is that this approach works significantly better than the phrase translation using the NMT system. Using the proposed approach, we obtained a Russian-English dictionary of lexical expressions, which can be used both as a ready-made dictionary and as a raw resource for manual dictionary construction. The resulting Russian-English phrase dictionary was placed on the Internet as a linguistic resource.

✉ Khusainova Albina M.
e-mail: a.khusainova@innopolis.ru

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

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

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

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

3. *Brown P. F., Pietra V. J. D., Pietra S. A. D. and Mercer R. L.* (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. Comput. Linguist. 1993 Jun. 19. P. 263–311.

4. *Och F. J. and Ney H.* (2003) A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 2003 Mar. 29. P. 19–51. – DOI: 10.1162/089120103321337421.

5. *Garcia M., García-Salido M. and Alonso-Ramos M.* (2019) Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics.

6. *Smadja F., McKeown K. and Hatzivassiloglou V.* (1996) Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Comput. Linguistics*. 22. P. 1–38.

7. *Kupiec J.* (1993) An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In 31st Annual Meeting of the Association for Computational Linguistics; 1993 Jun; *Columbus: Association for Computational Linguistics.* P. 17–22. DOI: 10.3115/981574.981577.

8. *Rivera O. M., Mitkov R. and Corpas Pastor G.* (2013) A flexible framework for collocation retrieval and translation from parallel and comparable corpora. *In Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technologies*. 2013 Sep. Nice.

9. *Seretan V. and Wehrli É.* (2007) Collocation translation based on sentence alignment and parsing. In Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs; 2007 Jun; Toulouse: ATALA. P. 375–384.

10. *Zenkel T., Wuebker J. and DeNero J.* (2020) End-to-End Neural Word Alignment Outperforms GIZA++. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020 Jul; Online: Association for Computational Linguistics. P. 1605–1617. – DOI: 10.18653/v1/2020.acl-main.146.

11. *Chen Y., Liu Y., Chen G., Jiang X. and Liu Q.* (2020) Accurate Word Alignment Induction from Neural Machine Translation. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2020 Nov; Online: Association for Computational Linguistics. P. 566–576. – DOI: 10.18653/v1/2020.emnlp-main.42.

12. *Koehn P., Axelrod A., Birch A., Callison-Burch C., Osborne M. and Talbot D.* (2005) Edinburgh system description for the 2005 IWSLT speech translation evaluation. *International Workshop on Spoken Language Translation*. 2005 Jan.

13. *Richardson J., Nakazawa T. and Kurohashi S.* (2014) Bilingual Dictionary Construction with Transliteration Filtering. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*; 2014 May; Reykjavik: European Language Resources Association (ELRA). P. 1013–1017.

14. *Daiga Deksne A. V.* (2018) A Workflow for Supplementing a Latvian-English Dictionary with Data from Parallel Corpora and a Reversed English-Latvian Dictionary. *In Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*; 2018 Jul; Ljubljana: Ljubljana University Press, Faculty of Arts. P. 127-135.

15. *Chen Y. J., Yang C. Y. H. and Chang J. S.* Improving Phrase Translation Based on Sentence Alignment of Chinese-English Parallel Corpus. In Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020); 2020 Sep; Taipei: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). p. 6–7.

16. *Schwenk H., Wenzek G., Edunov S., Grave E., Joulin A. and Fan A.* (2021) CCMatrix: Mining

Billions of High-Quality Parallel Sentences on the Web. *In 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); 2021 Aug; Online: Association for Computational Linguistics. P. 6490–6500. – DOI: 10.18653/v1/2021.acl-long.507.

17. *Tiedemann J.* (2012) Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12); 2012 May; Istanbul: European Language Resources Association (ELRA). P. 2214–2218.

18. *Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N. [et al.]* (2007) Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions; 2007 Jun; Prague: Association for Computational Linguistics. P. 177–180.

19. Moses. – Available at: https://www.statmt.org/moses/. (accessed: 06.11.2022).

20. Corpus dictionary of multiword lexical units (expressions). – Available at: https://ruscorpora.ru/page/obgrams/. (accessed: 06.11.2022).

21. Russian-english dictionary of collocations and phrases. – Available at: https://audio-class.ru/english-collocations/vocabulary-02.php. (accessed: 06.11.2022).

22. *Tiedemann J. and Thottingal S.* (2020) OPUS-MT – Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation; 2020 Nov; Lisboa: European Association for Machine Translation. P. 479–480.

23. MarianMT. – Available at: https://huggingface.co/docs/transformers/model_doc/marian. (accessed: 06.11.2022).

24. Helsinki-NLP/opus-mt-ru-en. – Available at: https://huggingface.co/Helsinki-NLP/opus-mt-ru-en. (accessed: 06.11.2022).

25. *Pecina P.* (2005) An Extensive Empirical Study of Collocation Extraction Methods. *In Proceedings of the ACL Student Research Workshop*; 2005 Jun; Ann: Association for Computational Linguistics. p. 13–18.

26. *Bhalla V. and Klimcikova K.* (2019) Evaluation of automatic collocation extraction methods for language learning. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications; 2019 Aug; Florence: Association for Computational Linguistics. P. 264–274. – DOI: 10.18653/v1/W19-4428.

Khusainova Albina M. — 4th year post-graduate student, assistant in Machine Learning and Knowledge Representation Laboratory, Innopolis University.
E-mail: a.khusainova@innopolis.ru
ORCID iD: https://orcid.org/0000-0002-0636-3449

Romanov Vitaly A. — 4th year post-graduate student, assistant in Industrial Software Production Laboratory, Innopolis University.
E-mail: v.romanov@innopolis.ru
ORCID iD: https://orcid.org/0000-0003-3772-0039

Khan Adil M. — Candidate of Science in Physics and Mathematics, Professor, Head of the Machine Learning and Knowledge Representation Laboratory, Innopolis University.
E-mail: a.khan@innopolis.ru
ORCID iD: https://orcid.org/0000-0003-2220-8518