ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2024, № 4, с. 121–128

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 519.711.2

УПОРЯДОЧИВАНИЕ ГИПОТЕЗ В МОДЕЛЯХ ПЕРЕВОДА С ИСПОЛЬЗОВАНИЕМ ЧЕЛОВЕЧЕСКОЙ РАЗМЕТКИ

© 2024 г. К. В. Воронцов[a], *, Н. А. Скачков[a], **

[a] *ВЦ ФИЦ ИУ РАН, Москва, Россия*

**e-mail: vokov@forecsys.ru*

***e-mail: nikolaj-skachkov@ya.ru*

Поступила в редакцию 31.01.2024 г.

После доработки 23.04.2024 г.

Принята к публикации 22.07.2024 г.

Современные системы машинного перевода обучаются на больших объемах параллельных данных, полученных с помощью эвристических методов обхода интернета. Низкое качество этих данных приводит к систематическим ошибкам перевода, которые могут быть достаточно заметными для человека. Для борьбы с такими ошибками предлагается интегрирование человеческих оценок гипотез переводной модели в процесс обучения системы перевода. Показано, что использование человеческих разметок позволяет не только вырастить общее качество перевода, но и заметно снизить количество систематических ошибок перевода. Кроме того, относительная простота человеческой разметки и ее применения для улучшения качества модели открывает новые возможности в области доменной адаптации моделей перевода под новые домены, что удалось показать на примере переводов заголовков товаров из интернет-магазинов.

Ключевые слова: машинный перевод, нейронная сеть, стохастический градиентный спуск, контрастное обучение, дообучение модели, обучение с негативными примерами.

DOI: 10.31857/S0002338824040074 **EDN:** UEFMST

HYPOTHESES RE-RANKING IN TRANSLATION MODELS USING HUMAN MARKUP

K. V. Vorontsov[a], *, N. A. Skachkov[a], **

[a] *Moscow, CC FRC CSC RAS*

**vokov@forecsys.ru*

***nikolaj-skachkov@ya.ru*

Modern machine translation systems are trained on large volumes of parallel data obtained using heuristic methods of the Internet bypassing. The poor quality of the data leads to systematic translation errors, which can be quite noticeable from the human point of view. To fix such errors a human based models hypotheses re-ranking is introduced in this work. In this paper the use of human markup is shown not only to increase the overall quality of translation, but also to significantly reduce the number of systematic translation errors. In addition, the relative simplicity of human markup and its integration in the model training process opens up new opportunities in the field of domain adaptation of translation models for new domains like online retail.

Keywords: Machine translation, neural network, stochastic gradient descent, contrastive learning, model fine-tuning, negative examples.

Введение. Создание систем автоматического перевода является одной из сложных задач анализа текстов естественного языка. Обучение алгоритмов, лежащих в основе таких систем, требует большого количества параллельных текстов на разных языках и существенно зависит от качества этих данных и степени их выравненности. Ввиду высокой стоимости работы профессиональных переводчиков данные для обучения алгоритмов перевода собираются автоматически с помощью эвристических алгоритмов. При этом высокая степень выравненности

отдельных обучающих примеров не гарантируется, что позволяет собирать большие объемы параллельных текстов. [1]

На собранных параллельных данных обучаются нейросетевые модели [2, 3], которые учатся восстанавливать каждое слово перевода при условии входного текста и предыдущих слов. Обученные таким образом модели, при должном размере модели, способны хорошо восстанавливать языковые закономерности и генерировать достаточно гладкие с человеческой точки зрения переводы. Однако качество перевода естественным образом связано с объемом и качеством собранных параллельных данных. Так из-за невозможности находить в интернете больших объемов хорошо выравненных переводов модели перевода страдают такими типовыми ошибками, как недостаточные переводы [4].

Для борьбы с типовыми ошибками существует способ интеграции негативных примеров в процесс обучения [4]. При использовании обучения с негативными примерами возникает возможность передать сети информацию, какие переводы являются неприемлемыми с точки зрения качества. В работе [4] предлагалось портить переводы из обучающего корпуса с помощью выкидывания случайных слов. В результате такого обучения модель перевода становится более внимательной к информации, содержащейся во входном предложении, и количество ошибок с пропуском слов в переводе уменьшается. Однако генерация негативных примеров по простым шаблонам несет в себе некоторые ограничения. Так, негативные примеры, полученные с помощью генерации по шаблону, могут оказаться слишком простыми и не покрывать все возможные случаи внутри одного класса ошибок. Кроме того, для каждого отдельного класса ошибок требуется описывать свой шаблон для генерации негативных примеров. Если для недостаточных переводов может подойти выкидывание случайных слов из правильного перевода, то для таких ошибок как неверное согласование текста с точки зрения языка выбрать правильный шаблон становится сложнее. Сама по себе необходимость придумывать шаблоны для различных видов ошибок делает процесс улучшения качества машинного перевода более трудоемким и менее масштабируемым.

К проблеме выбора негативных примеров можно подойти с другой стороны. Для каждого текста на входном языке можно получить несколько переводов модели, например с помощью процедуры разнообразного поиска в ширину [5] или сэмплирования с температурой. При этом в разнообразных переводах могут случайным образом содержаться или не содержаться типовые ошибки данной модели. Тогда, если выбрать лучший по качеству перевод среди имеющихся, то его можно использовать в качестве позитивного примера, а все остальные относительно него будут негативными. Каждый негативный пример, полученный таким образом, оказывается сложным с точки зрения модели, так как он оказался достаточно вероятным, чтобы она его сгенерировала. Соответственно такое обучение представляет из себя упорядочивание гипотез модели с точки зрения некой метрики качества.

Остается вопрос: как выбрать метрику качества для ранжирования сгенерированных моделью переводов? В работе представлен подход к обучению моделей машинного перевода с помощью упорядочивания гипотез модели на основе человеческих оценок. Показано, что применение данного подхода позволяет не только улучшить качество перевода, но и заметно снизить долю типовых ошибок перевода, которыми модель перевода изначально страдала. Кроме того, благодаря тому, что модель перевода в подходе не учится с нуля, дообучение с упорядочиванием гипотез не требует большого количества данных человеческой разметки.

Более того, обучение с упорядочиванием гипотез на основе человеческих оценок не требует наличия параллельных данных или эталонных переводов. Это открывает возможность для улучшения качества модели перевода на тех доменах, где нет данных для доменной адаптации. В статье исследуется применимость предложенной процедуры для улучшения качества переводов заголовков товаров из домена электронной коммерции. Этим текстам свойственна специфичная структура текстов и лексика, что является причиной их сложности для систем автоматического перевода.

1. Постановка задачи. Теперь рассмотрим вероятностные модели, лежащие в основе систем автоматического перевода текстов естественного языка, а также опишем подход с упорядочиванием гипотез перевода на базе человеческих оценок.

1.1. Постановка задачи машинного перевода. Перейдем к математической постановке задачи машинного перевода. Пусть задано множество параллельных данных, состоящее из пар текстов $\{(x_i. y_i)\}_{i=1}^{N}$. Тексты x_i написаны на языке входа, а тексты y_i — на целевом языке и являются переводами соответствующих текстов x_i. Тогда обучение переводной модели будет заключаться в максимизации правдоподобия переводов при условии входных текстов

для всех объектов выборки. Основываясь на методе максимального правдоподобия и переходя к логарифму правдоподобия, получим

$$\sum_{i=1}^{N} \log P_\theta\left(y_i \mid x_i\right) \to \max_\theta, \tag{1.1}$$

где θ — параметры обучаемой переводной модели, а $P_\theta(\cdot)$ — функция правдоподобия модели. Обученная с помощью максимизации правдоподобия (1.1) модель умеет оценивать вероятность перевод y для входного x для любого y. Однако для модели перевода этого недостаточно, так как она должна быть способна генерировать переводы, а не только оценивать их вероятность. Для генерации перевода с помощью такой оценивающей модели необходимо перебрать все возможные тексты на выходном языке и, оценив каждый из них с помощью модели, выбрать наиболее вероятный перевод. Так как число текстов на выходном языке бесконечно, генерация с помощью такой модели невозможна. Для решения этой проблемы используют авторегрессионные модели перевода [2]. В данном подходе вводится дополнительное ограничение, что t-е слово перевода зависит только от предыдущих слов. Правдоподобие модели перевода с этим ограничением записывается следующим образом:

$$\log P_\theta\left(y \mid x\right) = \sum_{t=1}^{|y|} \log P_\theta\left(y^t \mid y^{<t}, x\right), \tag{1.2}$$

где y^t — t-е слово перевода, $y^{<t}$ – префикс перевода для t-го слова, а $P_\theta(y^t \mid y^{<t}, x)$ — моделируемая вероятность t-го слова перевода при условии префикса и входного текста.

Для обученной с авторегрессионной функцией потерь (1.2) модели генерация перевода может осуществляться при помощи выбора наиболее вероятного слова y^t.

1.2. К о н т р а с т н о е о б у ч е н и е с н е г а т и в н ы м и п р и м е р а м и. Для борьбы с систематическими ошибками перевода, которые возникают из-за низкого качества выравнивания в параллельных данных, в работе [4] предложен подход по обучению с негативными примерами. Для генерации негативных примеров, т.е. заведомо неправильных переводов, переводы портятся по некоторому шаблону. Так, для параллельной пары (x, y) строится испорченный перевод $y\_$, полученный из y с помощью шаблона t. Пример y далее будем называть положительным примером, а испорченный перевод $y\_$ — отрицательным. Для борьбы с ошибками пропуска слов в переводе авторы предлагают выбрасывать случайные слова в переводе. Тогда шаблон t можно описать следующим образом:

$$t(y) = y_1 .. y_{t-1} y_{t+1} \cdots y_{|y|}, \quad t \sim \mathcal{U}\left[1, \ldots, |y|\right],$$

где y состоит из слов $y_1 \ldots y_{|y|}$, $|y|$ — длина текста, а t выбирается случайно из целых чисел от 1 до $|y|$.

Для интеграции в процесс обучения негативных примеров авторы после обучения с авторегрессионной функции потерь (1.2) использовали дообучение модели с контрастной функцией потерь:

$$L_\alpha\left(x, y, y\_, \theta\right) = \max\left(0, \log P_\theta\left(y\_ \mid x\right) - \log P_0\left(y \mid x\right) + \alpha\right), \quad y\_ = t\left(y\right), \tag{1.3}$$

при которой увеличивается вероятность положительного перевода y текста x относительного более плохого перевода $y\_$. Таким образом, при обучении с контрастной функцией потерь (1.3) модель учится правильно упорядочивать положительный и отрицательный примеры. Для тех обучающих примеров, у которых положительный пример более вероятен, чем отрицательный на значение, превышающее отступ α, контрастная функция потерь (1.3) обращается в ноль и эти примеры не участвуют в обучении.

Можно заметить, что при дообучении с контрастной функцией потерь (1.3) модель может «забыть» задачу авторегрессионной генерации и, следовательно, потерять возможность итеративно генерировать перевод. Это объясняется тем, что контрастная функция потерь (1.3) действует только на уровне предложения и воздействие на пословные предсказания $P_\theta(y_t \mid y_{<t}, x)$, из которых складывается вероятность всего перевода, может быть достаточно шумным и непредсказуемым. Для исключения возможности возникновения такой проблемы в данной работе дообучение происходит с функцией потерь, являющейся линейной комбинацией авторегрессионной функции потерь (1.2) и контрастной функции потерь (1.3):

$$L_{\alpha,\beta} = \beta \log P_\theta\left(y \mid x\right) + \max\left(0, \log P_\theta\left(y\_ \mid x\right) - \log P_\theta\left(y \mid x\right) + \alpha\right). \tag{1.4}$$

Подбор параметров α и β описан в разд. 2.4, где будет показано, что дообучение только на контрастную функцию потерь (1.3) действительно приводит к более плохому результату, чем

обучение на линейную комбинацию (1.4). Далее обучение с функцией потерь (1.4) будем называть контрастным обучением.

1.3. Выбор лучшего перевода с помощью человеческой разметки. Вместо генерации негативных примеров по шаблону в работе предлагается генерировать несколько переводов из одной модели с помощью разнообразного поиска в ширину [5] и упорядочивать их по качеству с помощью человеческой разметки. Как уже было описано, таким образом модель получает в качестве негативных примеров достаточно вероятные с точки зрения этой же модели переводы. Благодаря этому факту полученные негативные примеры по построению будут достаточно сложными для модели.

Для нахождения лучшего перевода среди двух переводов модели необходимо разработать инструкцию и шаблон разметки, чтобы унифицировать представления о задании у размечающих. В используемом задании предлагалось для представленного текста на входном языке выбрать один из двух переводов в качестве лучшего либо указать, что представленные переводы одинакового качества. Кроме того, для упрощения разметки различия в переводах подсвечивались. Пример задания разметки можно увидеть на рисунке. В инструкции к заданию указывались различные критерии оценки, такие, как точность передаваемого смысла, грамматическая корректность перевода, правильность расстановки пунктуации и выбора капитализации. Указанные категории ошибок в инструкции не упорядочивались по грубости, а лишь указывались в качестве напоминания. Это было сделано для того, чтобы оставить свободу выбора более грубой категории ошибок самому размечающему. Результатом такой человеческой разметки считаем упорядоченные пары переводов исходного текста, в каждой паре размечающий указывает, какой перевод лучше.

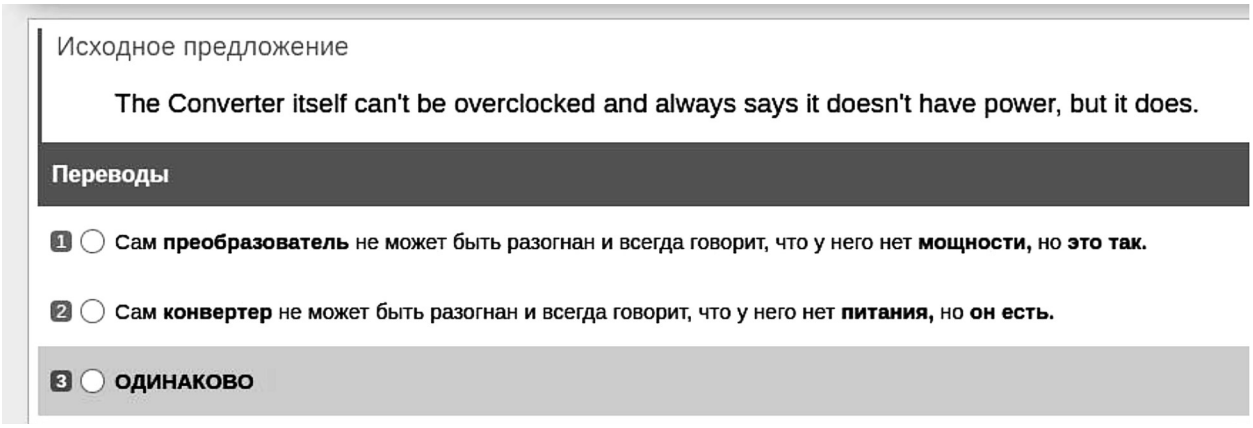

Рис. 1. Пример задания для человеческих разметчиков

Для уменьшения количества некачественной разметки был составлен экзамен для допуска к разметке. Во время разметки задания размечающим периодически показывались вопросы с заготовленным ответом. При ошибках на таких вопросах размечающие не допускались к продолжению разметки. Последнее помогало бороться с усталостью размечающих в процессе оценки переводов и потерей внимательности.

1.4. Оценка качества перевода. Для оценки качества переводов дообученных моделей будем использовать автоматическую метрику BLEU [6]. Эта метрика требует наличия эталонных переводов для тестового корпуса и показывает достаточно высокую корреляцию с оценками людей. BLEU рассчитывается на основе пересечения n-грамм в автоматическом и эталонном переводах одного предложения. Для каждой n-граммы длины n рассчитывается P_n – отношение частоты n-граммы среди всех кандидатов к частоте n-граммы среди всех эталонных переводов. При этом частота в кандидате ограничивается значением в эталоне, чтобы отношение частот было ограничено сверху единицей:

$$P_n = \frac{\sum\limits_{\text{n-gram}\in C} \text{Count}_{\text{clip}}\left(\text{n-gram}\right)}{\sum\limits_{\text{n-gram}'\in C'} \text{Count}\left(\text{n-gram}'\right)},$$

где C — переводы тестового корпуса моделью, C' — эталонные переводы тестового корпуса, а $\text{Count}_{\text{clip}}$ — частота n-граммы в переводе, ограниченное частотой этой же n-граммы в эталоне. Далее сама метрика BLEU считается как геометрическое среднее значений P_n, умноженное на константу BP (brevity penalty). Константа BP и нормализация в формуле BLEU при этом выбираются эмпирически:

$$\text{BP} = \min\left(\exp(1 - r/c), 1\right),$$

$$\text{BLEU} = \text{BP} \sum_{n=1}^{4} \frac{1}{n} \log P_n,$$

где r — суммарная длина эталонных переводов тестового корпуса; c' — суммарная длина переводов тестового корпуса моделью. Умножение на константу BP предлагается авторами метрики для уменьшения штрафа для более коротких предложений. Это необходимо из-за того, что более длинные переводы в среднем содержат больше случайных пересечений по n-граммам с эталонными текстами.

При проведении экспериментов метрика BLEU вычисляется с эталонными переводами, подготовленными профессиональными переводчиками. Для экспериментов использовались тестовые корпуса для русско-английского направления, подготовленные к конференции WMT- 2019 [7]. Размер тестового корпуса составляет 3000 предложений, исходные предложения выбирались из новостных статей.

Кроме автоматической метрики BLEU в данной работе для сравнения обученных моделей перевода будет применяться человеческая разметка на подобии описанной в разд. 1.3 (рисунок). Чтобы снизить эффект от переобучения под предпочтения размечающих, в оценке переводов будут использоваться только разметчики, не участвовавшие в разметке обучающих данных. Оценка на основе людей необходима для того, чтобы количественно оценить изменения в модели, а также получить возможность оценивать качество перевода текстов, не имеющих эталонных переводов.

Кроме сравнения переводов различных моделей, большой интерес представляет то, как предложенная процедура дообучения с упорядочиванием гипотез на основе человеческой разметки уменьшает долю систематических ошибок перевода. Для оценки этого эффекта часть переводов показывалась размечающему с вопросом «является ли перевод недостаточным».

2. Эксперименты. Перейдем к описанию экспериментов и условий их проведения. Кроме того, приведем основные результаты, полученные при контрастном дообучении модели на человеческую разметку (1.4).

2.1. Архитектура модели. Обучаемые в данной работе модели перевода основаны на архитектуре Transformer [3]. Данная архитектура подразумевает, что модель состоит из кодировщика и декодировщика, каждый из которых в свою очередь состоит из нескольких блоков одинаковой структуры, выполняющихся последовательно.

Каждый блок имеет свой набор параметров и применяется к векторным представлениям слов предложения, полученных от предыдущего блока, и возвращает новые обогащенные векторные представления слов. Внутри самого блока происходит обогащение векторного представления контекстной информацией с помощью механизма внимания, а также к векторному представлению применяется нелинейное преобразование. В блоках декодировщика кроме контекстной информации векторные представления обогащаются еще и информацией о входном предложении с помощью механизма внимания на выход кодировщика. При этом контекстная информация в декодировщике ограничивается только левым контекстом. Такая архитектура позволяет осуществлять итеративную генерацию для моделей, обученных с авторегрессионной функцией потерь (1.2).

2.2. Условия экспериментов. В экспериментах проводится дообучение модели с различными функциями потерь и с помощью размеченных человеком переводов. В качестве предобученной модели в экспериментах использовалась модель из библиотеки fairseq, являющаяся победителем на направлении с русского на английский WMT-2019 [7]. Модель соответствует архитектуре Transformer-big и обладает размерностью векторных представлений:

1024 для внутренних представлений,

4096 для представлений внутри FFN,

16 голов внимания.

Дообучение модели производилось с помощью оптимизатора Adam [8] с нагревом в течение 1000 шагов и охлаждением по sqrt-расписанию в течение оставшихся 4000 шагов. Обучались модели на сервере с 4 GPU Tesla M40.

2.3. Данные для обучения. Эксперименты, как уже было описано, проводятся на направлении с русского языка на английский. Предобученная модель обучалась на данных, предоставленных для соревнования WMT-2019 [7], которые состоят из данных Paracrawl v3, Common Crawl, News Commentary и других корпусов. Для дообучения используются данные разметки, осуществленной на основе переводов текстов из датасета News Commentary.

Для разметки выбраны случайно 10000 текстов из указанного обучающего набора. Для каждого из них составлены по два перевода предобученной модели с помощью процедуры разнообразного поиска в ширину [5]. Далее тексты, у которы длина входного текста и перевода не превышает три слова, а также перевод содержит больше половины английских слов, отфильтровываются. Из оставшихся переводов выбирается лучший на основе процедуры, указанной в разд. 1.3.

Из полученных данных разметки выбираются только те примеры, где размечающие выбрали какой-либо перевод в качестве лучшего. Оказалось, что из всех обучающих примеров в 33% случаев по оценке человека перевод, который был менее вероятен с точки зрения модели, оказывался лучше, чем более вероятный. В 15% случаев с точки зрения размечающих качество оказывалось одинаковым. Эти примеры не используются для дообучения моделей, так как их не удалось упорядочить по качеству.

2.4. Эксперименты с контрастной функцией потерь. Для выбора гиперпараметров отступа α и веса β в контрастной функции потерь с человеческой разметкой (1.4), а также гиперпараметра отступа α в контрастном обучении с сгенерированными по шаблону негативными примерами (1.3) были обучены модели с перебором гиперпараметров по сеткам. При отборе моделей осуществлялся выбор лучшей конфигурации по метрике BLEU на датасете WMT-17 с более старого соревнования по машинному переводу. Отбор моделей по WMT-19 не проводился, чтобы избежать переобучения под тестовую выборку.

Для контрастного обучения с человеческой разметкой (1.4) оптимальные значения гиперпараметров оказались $\alpha = 0.3$, $\beta = 0.1$. Для обучения с негативными примерами, сгенерированными по шаблону (1.3), оптимальное значение α оказалось равным 1.0.

3. Сравнение подходов. Теперь обучим модели, представляющие описанные подходы к обучению и дообучению машинного перевода, и оценим их качество. Наибольший интерес представляют следующие модели:

базовая – модель, взятая из библиотеки fairseq, которая дообучалась в остальных экспериментах;

дообученная на параллельные данные – модель, которая училась столько же шагов, сколько и остальные дообученные модели с авторегрессионной функцией потерь (1.2);

дообученная на победителя разметки – модель, которая дообучалась с авторегрессионной функцией потерь (1.2) на тот перевод, который победил в разметке;

дообученная с шаблонными негативными примерами – модель, которая дообучалась с контрастной функцией потерь (1.3) с негативными примерами, которые генерировались по шаблону;

дообученная на человеческую разметку – модель, которая обучалась с контрастной функцией потерь (1.4) на упорядоченные по качеству с помощью человеческой разметки переводы.

Результаты оценки обученных моделей по BLEU на тестовом наборе WMT-19 можно увидеть в табл. 1. Как можно заметить, что обе модели, которые дообучались на данные, полученные с помощью человеческой разметки на качество, имеют значительный прирост на тестовом наборе. Причем модель, которая дообучалась с контрастной функцией потерь (1.4), превосходит по BLEU модель, обучавшуюся только на победителя без негативных примеров. Отдельный интерес представляет вопрос: насколько обучение с человеческой разметкой помогает справиться с систематическими ошибками перевода? Размечая переводы моделей на предмет того, является ли перевод недостаточным, т.е. в нем отсутствуют какие-то части, представленные во входном предложении, удалось выяснить, что дообучение на человеческую разметку наиболее полно решает данную проблему. Это можно объяснить тем, что негативные примеры, полученные из самой же модели, оказываются существенно более сложными, чем негативные примеры, сгенерированные по шаблону. Так, доля недостаточных переводов при контрастном дообучении с человеческой разметкой (1.4) падает с 4 до 1%, тогда как дообучение с шаблонными негативными примерами (1.3) понижает долю таких ошибок лишь до 2%.

Рассмотрим теперь, что происходит с качеством перевода с точки зрения человеческих оценок. Для этого разметим и сравним переводы разных моделей с помощью разметки, описанной в разд. 1.3. При этом размечающие выбираются только те, кто не участвовал в разметке обучающих данных. В итоге получается, что контрастное дообучение на человеческую разметку (1.4) улучшает перевод в 15% случаев, тогда как обучение с шаблонными негативами улучшает перевод лишь в 3% случаев. Примеры, где дообучение с человеческой разметкой улучшает перевод с точки зрения размечающих, можно увидеть в табл. 2.

Таблица 1. Сравнение дообученных моделей на тестовом наборе WMT19 на направлении с английского на русский

Модель	Функция потерь	BLEU	Недостаточные переводы, %
Базовая	Авторегрессионная (1.2)	35.6	4
Дообучение на переводные данные	Авторегрессионная (1.2)	35.7	–
Дообучение на победителя разметки	Авторегрессионная (1.2)	36.7	–
Дообучение с шаблонными негативами	Контрастная (1.3)	35.8	2
Дообучение на разметку	Контрастная (1.4)	**37.3**	1

Таблица 2. Примеры переводов моделей.

Вход:	а в одиночку как-то скучно
Базовая:	and alone as it is boring
Дообученная:	and it’s kind of boring alone
Вход:	все время загорается красным
Базовая:	it’s always red
Дообученная:	it lights up red all the time
Вход:	у меня немного опыта путешествий
Базовая:	I don’t have much experience
Дообученная:	I have a little experience in traveling

Входом обозначены тексты, подаваемые на вход моделям. Базовой называется модель до дообучения на разметку (1.4). Дообученной обозначена модель после контрастного дообучения на человеческую разметку (1.4)

3.1. Эксперименты с доменной адаптацией. Рассмотрим теперь применимость подхода с контрастным обучением на человеческую разметку (1.4) к задаче доменной адаптации. Как уже было сказано, при использовании человеческой разметки нет необходимости в параллельных данных, что открывает возможности по улучшению качества перевода на тех доменах, где сложно найти параллельные данные достаточно хорошего качества.

Для экспериментов было выбрано направление перевода с английского на русский на домене заголовков товаров из интернет-магазинов. В качестве предобученной модели использовалась все так же обученная с авторегрессионной функцией потерь (1.2) модель перевода. Для доменной адаптации были размечены переводы моделью из 10000 заголовков с помощью процедуры описанной в разд. 1.3.

Стоит заметить, что заголовки товаров представляют из себя достаточно сложный домен для машинного перевода. Это связано с тем, что такие тексты обладают нестандартной структурой: в них зачастую отсутствует сказуемое, а также присутствует большое количество определений и перечислений. В табл. 3 можно увидеть пример того, насколько плохо справляется модель с переводом входного текста до процедуры дообучения.

Таблица 3. Пример переводов моделей при доменной адаптации.

Вход:	car temporary parking card luminous calling phone number cards with sucker plate
Базовая:	автомобильная временная парковочная карта, светящиеся карточки с номером телефона для звонков с присоской
Дообученная:	светящиеся карточки с номерами телефонов для временной парковки автомобилей с присоской

Входом обозначены тексты, подаваемые на вход моделям. Базовой называется модель до дообучения на разметку (1.4). Дообученной обозначена модель после контрастного дообучения на человеческую разметку (1.4)

Так как для данного домена отсутствуют качественные тестовые наборы, оценка качества проводилась с помощью разметки. По итогам разметки переводов моделями из 1000 заголовков товаров получилось, что качество после дообучения на контрастную функцию потерь с человеческой разметкой (1.4) выросло на 40% заголовков. В табл. 3 можно также увидеть то, как улучшился перевод заголовков после процедуры дообучения. В среднем стоит отметить, что после дообучения переводы стали более гладкими с точки зрения русского языка и части заголовков стали переводиться согласованно друг с другом.

Заключение. Проведено исследование по возможности использования человеческой разметки для улучшения качества машинного перевода. Удалось показать, что упорядочивание переводов модели по качеству с помощью разметки позволяет заметно усилить эффект от обучения перевода с негативными примерами. Благодаря сложности получаемых из разметки примеров модель после дообучения показывает более высокие результаты как по тестовым наборам, как и с точки зрения человеческих оценок. Также данная процедура толкает модель исправлять свои же систематические ошибки, доля таких ошибок, как недостаточные переводы, заметно падает. Более того, модель, обученная с негативными примерами, сгенерированными по шаблону специально для борьбы с данным типом ошибок, чаще допускает недостаточные переводы. Данный эффект объясняется тем, что шаблон не способен описать достаточно сложные негативные примеры и модель, учащаяся на разметке своих же ошибок, исправляет их лучше.

Кроме того, предложенная процедура открывает возможности для более эффективной доменной адаптации. Для тех доменов, где есть недостаток качественных параллельных данных, вместо привлечения профессиональных переводчиков появляется возможность улучшения качества с помощью разметки переводов модели. Так, на домене заголовков товаров из интернет-магазинов с помощью дообучения модели на разметку переводов удалось заметно поднять качество модели с точки зрения человеческих оценок.

СПИСОК ЛИТЕРАТУРЫ

1. *Bañón M., Chen P., Haddow B. et. al.* ParaCrawl: Web-Scale Acquisition of Parallel Corpora // Proc. 58th Annual Meeting of the Association for Computational Linguistics. Seattle, 2020. P. 4555–4567.
2. *Stahlberg F.* Neural Machine Translation: A Review // J. Artific. Intelligence Res. 2020. № 69. P. 343–418.
3. *Vaswani A., Shazeer N., Parmar N. et. al.* Attention is All You Need // Proc. 31st Intern. Conf. on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook. N.Y., 2017. P. 6000–6010.
4. *Yang Z., Cheng Y., Liu Y. et. al.* Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach // Proc. 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019. P. 6191–6196.
5. *Vijayakumar A.K., Cogswell M., Selvaraju R.R. et. al.* Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models // ArXiv. 2016. abs/1610.02424.
6. *Papineni K., Roukos S., Ward T. et. al.* Bleu: a Method for Automatic Evaluation of Machine Translation // Proc. 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, 2002. P. 311–318.
7. *Barrault L., Bojar O.R., Costa-jussà M. et al.* Findings of the Conf. on Machine Translation (WMT19) // Proc. Fourth Conf. on Machine Translation. Florence, 2019. V. 2: Shared Task Papers.
8. *Kingma D.P., Ba J.* Adam: A Method for Stochastic Optimization // 3rd Intern. Conf. on Learning Representations (ICLR). San Diego, CA, 2015.