УДК 004.421, 004.912 https://doi.org/10.51301/vest.su.2021.v143.i1.08

Д.Р. Рахимова, Н.М. Пазылхан*, А.А. Кульжанова, Ж.Г. Ален
Казахский национальный университет им. Аль-Фараби, Алматы, Казахстан
*e-mail: npazylhan@gmail.com

РАЗРАБОТКА МОДЕЛИ И ПРОГРАММНОГО РЕШЕНИЯ ЗАДАЧИ ОПРЕДЕЛЕНИЯ НЕИЗВЕСТНЫХ СЛОВ ПРИ ПОСТРЕДАКТИРОВАНИИ МАШИННОГО ПЕРЕВОДА

Аннотация. Машинный перевод - это технология последовательного перевода текстов с одного языка на другой компьютерной программой. В результате машинного перевода всегда есть определенные недостатки, которые можно решить постредактированием. Постредактирование - человеческая обработка текста после машинного перевода. Сегодня многие поставщики языковых услуг активно развивают это направление, разрабатывая методы обучения редакторов и методы постредактирования. В работе представлен обзор современных проблем систем машинного перевода. В данной работе рассматривается задача определения неизвестных слов при постредактировании машинного перевода для казахского языка. Проведен анализ существующих методов нахождения неизвестных слов при постредактировании машинного перевода. Представлены модель определения неизвестных слов при постредактировании машинного перевода для англо-казахского и русско-казахского языка, ппрактические результаты и программная реализация.

Ключевые слова: машинный перевод, NLTK, мморфологический анализ, неизвестные слова, постредактирование машинного перевода.

Обзор существующих методов. Каждая система машинного перевода сталкивается с проблемой неизвестных слов. Доступные в настоящее время корпуса, особенно для языков с меньшими ресурсами, не охватывают все возможные слова на данном языке, и часто добавляются новые слова. Словарь - неотъемлемая часть любой системы машинного перевода. «Неизвестный» определяется как слово, для которого нет словарной статьи. Проблема неизвестного слова особенно серьезна для небольших портативных систем перевода, поскольку здесь словарный запас должен быть ограничен, чтобы можно было разместить систему перевода на портативном устройстве [1]. Имена (имена собственные), акронимы, аббревиатуры, терминология и слова, взятые из других языков (иностранные слова), являются одними из основных источников, которые вносят вклад в список неизвестных. Хотя делается попытка создать лексическую базу данных для конкретного домена, содержащую терминологию, аббревиатуры, акронимы и имена собственные, которые могут использоваться в этом домене, невозможно сделать ее включающей все. Однако грамматические правила построения рода, числа, номинализации глаголов или форм соответствуют таковым для используемого языка независимо от их происхождения. Эта вызывает частую встречу неизвестных слов в повседневное общение [2].

Текущие системы SMT либо отбрасывают неизвестные слова, либо буквально копируют их в вывод. Однако качество нейромашинного перевода пока не приближается к профессиональному переводу. Основная проблема нейронного машинного перевода - необходимость больших объемов параллельных вложений необходимые для обучения нейромашинному переводу. Это особенно актуально для малоресурсных языков, которым относится казахский язык. Способы решения этой проблемы: создание естественных параллельных корпусов профессиональными переводчиками или создание синтетических параллельных корпусов. Первый случай - процесс, требующий значительных ресурсов; Во втором случае возможны различные подходы к созданию синтетических параллельных пакетов. На качество нейронного машинного перевода также влияет проблема неизвестных

слов, то есть слов, находящихся за пределами словаря системы машинного перевода (Out Of Vocabulary - OOV) [3, 4].

Для решения проблемы неизвестных слов было предложено несколько подходов, которые можно разделить на три категории. Первая категория подходов направлена на повышение скорости расчета производительности инструмента softmax, чтобы он мог поддерживать очень большой словарный запас. Вторая категория использует информацию из контекста. В частности, применительно к задаче машинного перевода, система учится указывать некоторые слова в исходном предложении и копировать их в целевое предложение [6]. В этой работе авторы определяют новую методологию, которая устраняет это узкое место и предоставляет крупномасштабные контролируемые данные о понимании прочитанного. Это позволяет разработать класс глубоких нейронных сетей, основанных на внимании, которые учатся читать реальные документы и отвечать на сложные вопросы с минимальным предварительным знанием структуры языка. При настройке ответа на вопрос в контексте использовались заполнители для именованных объектов [7]. Третья категория подходов изменяет саму единицу ввода/вывода со слов на более низкое разрешение, такое как символы или байтовые коды [8, 9]. Главное преимущество этого подхода заключающееся в том, что может меньше страдать от проблемы неизвестных слов, обучение обычно становится намного более трудным, поскольку длина последовательностей значительно увеличивается [4].

При решении задачи дополнительного словарного запаса уделяется внимание тому, как правильно переводить дополнительный словарный запас. Для этого используются дополнительные ресурсы, такие как сопоставимые данные и тезаурус синонимов [10]. Заметным исключением является работа, в которой также уделяется внимание синтаксической и семантической роли слов вне словарного запаса и предлагается заменить слова вне словарного запаса с похожими словами во время тестирования [3, 11]. Следующем работе предложен и реализован эффективный метод решения проблемы неизвестных слов. Авторы предлагает решить проблему редких слов, обучив систему NMT отслеживать происхождение неизвестных слов в целевых предложениях. Если бы мы знали исходное слово, отвечающее за каждое неизвестное целевое слово, мы могли бы ввести этап постобработки, который заменил бы каждое UNK в выводе системы переводом исходного слова, используя словарь или перевод идентичности. Авторы обучили систему NMT на данных, которые были дополнены выходными данными алгоритма выравнивания слов, который позволял системе NMT отображать для каждого слова вне словаря в целевом предложении позицию соответствующего ему слова в исходном предложении. Эта информация была позже использована при постобработке фаза, которая переводит каждое слово вне словаря с помощью словаря [6].

В работе предлагается метод обработки редких и неизвестных слов для моделей нейронных сетей с использованием механизма внимания. Их модель использует два слоя softmax для предсказания следующего слова в моделях условного языка: один предсказывает местоположение слова в исходном предложении, а другой предсказывает слово в словаре краткого списка. На каждом временном шаге решение о том, какой слой softmax использовать, адаптивно принимает многослойный перцептрон, который зависит от контекста [12]. Для решения проблемы неизвестных слов предлагается метод замены-перевода-восстановления [13]. На этапе подстановки редкие слова в тестовом предложении заменяются по подобным словарным словам на основе модели подобия, полученной из одноязычных данных. На этапах перевода и восстановления предложение будет переведено с помощью модели, обученной новым двуязычным данным с заменой редких слов [4].

Как только слово объявляется неизвестным, оно проверяется на предмет возможного имени или аббревиатуры с помощью некоторых эвристических методов. Некоторые из распространенных эвристик, используемых для английского языка, заключаются в том, что

имена собственные начинаются с верхнего регистра, все акронимы в верхнем регистре. После этого начинается процесс определения типа неизвестного слово [2].

Описание модели задачи определения неизвестных слов при постредактировании машинного перевода. Описание модели задачи определения неизвестных слов при постредактировании машинного перевода. Определение типа неизвестного слово:

Система перевода на основе правил обычно состоит из следующих общих шагов:

- Морфологический анализ каждого слова в исходном предложении;
- Разбор входного исходного предложения на основе синтаксических категорий, полученных морфологическим анализатором;
- Преобразование дерева синтаксического анализа, полученного выше, в целевое дерево;
- Создание текста на целевом языке из преобразованного дерева;

Точно так же система перевода на основе примеров обычно состоит из следующих общих шагов:

- Морфологический анализ каждого слова в исходном предложении;
- Нахождение исходного предложения в базе примеров, имеющего минимальное расстояние с исходным предложением ввода, на основе используемых критериев расстояния. Это расстояние обычно вычисляется на основе синтаксической и семантической информации, полученной морфологическим анализатором и поиском по словарю и генерируя текст на целевом языке.

Неизвестное слово обнаруживается морфологическим анализатором. Морфологический анализатор пытается извлечь корневое слово в соответствии с грамматическими правилами [5]. Затем выполняется поиск данного слова и его корня в корпусе сделанное для казахского языка, известной терминологии и сокращении. Если запись найдена, ее синтаксическая категория и другая информация извлекается из корпуса или списков и отправляется для дальнейшей обработки. Возможно, что этот процесс может дать более одной категории и значения. В случае если соответствующая запись не найдена в словаре или известных списках, она объявляется неизвестной.Как только слово объявляется неизвестным, оно проверяется на предмет возможного имени или аббревиатуры с помощью некоторых эвристических методов. Некоторые из распространенных эвристик, используемых для английского языка, заключаются в том, что имена собственные начинаются с верхнего регистра, все акронимы в верхнем регистре. После этого начинается процесс определения типа неизвестного слово [2].В случае системы перевода, основанной на правилах, на этапе синтаксического анализа, если подходит более одной категории для неизвестного слова, используются дополнительные эвристики для выбора наиболее перспективной. Суффиксы типа «ing», «d», «ed» или «en» предполагают, что они могут быть производными от глагола. Точно так же неизвестное слово с суффиксом «ous» или «ly» предполагает, что это слово могло быть наречием. Если проблема все еще остается нерешенной, присваивается категория в соответствии с наиболее частым деревом синтаксического анализа. Определение категории неизвестного слова в случае подхода на основе примеров является простой задачей сопоставления и вычисления расстояния. Очевидно, потребуется большее количество поисков, поскольку поиск должен выполняться для каждой постулируемой категории неизвестное слово.

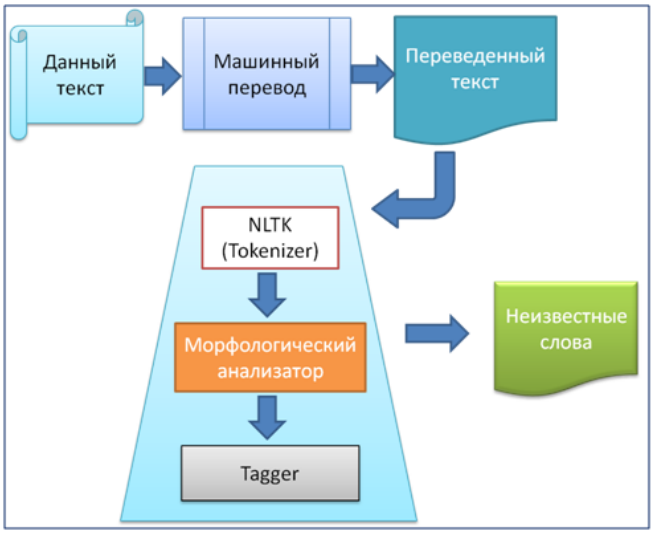

Рисунок 1. Модель задачи определения неизвестных слов при постредактировании машинного перевода

На рисунке 1 показано модель задачи определения неизвестных слов при постредактировании машинного перевода англо-казахского и русско-казахского языка. Алгоритм работы:

1. Дан исходный текст на английском или на русском языке.
2. Система машинного перевода переводит исходный текст.
3. Переведенный текст с помощью системы машинного перевода передается на NLTK(Tokenizer). С помощью NLTK(Tokenizer) разделим текст на отдельные слова.
4. После разделение предложений на отдельные слова эти слова обрабатывается с помощью морфологического анализатора.
5. Tagger анализирует отдельные слова, с помощью морфологического анализатора и пометить слова.
6. Необработанные неизвестные слова хранятся в отдельном файле.

С помощью этой модели улучшиться качество перевода, но не все слова найденные в корпусе могут гарантировать хороший перевод с контекстом неизвестного слова, поскольку им обычно не хватает соответствующей словарной записи. Таким образом, если найденное слово в корпусе объединение контекста неизвестного слова соответствует записи в таблице фраз, это облегчит лексический выбор и переупорядочение слов в окружающих словах.

Созданный морфологический анализатор был взять в основу морфологического анализатора сделанное в «*nlacslab*« [4]. Морфологический анализатор возвращает для данного слова следующий триплет: <*словарная форма*>, <*часть речи*> и <*грамматические характеристики*>, т.е. все возможные анализы. Морфологический анализатор, включенный в текущую работу, представляет собой реализацию анализатора, управляемого данными.

Поиск неизвестных слов в словарном запасе

По данным оценки и данных обучения мы можем легко отличить неизвестные слова от слов в словарном запасе. Предположим, что набор неизвестных слов - это UW, а набор слов в словаре - это IW. Для каждого неизвестного слова UW наша цель - найти наиболее подходящее слово IW* из IW, чтобы IW* имел наиболее похожую семантическую функцию с UW. С помощью функции подобия, определенной выше, можем использовать следующую формулу для достижения нашей цели:

$$IW^* = \mathbf{argmax}_{IW} \mathrm{Sim}(UW, IW) \qquad (1)$$

Однако мы обнаружили, что использование этой формулы без каких-либо ограничений обычно не дает хороших результатов. Следовательно, мы требуем, чтобы получившееся словарное слово IW* имело согласованную часть речи с неизвестным словом UW. Соответственно, формула поиска будет такой:

$$IW^* = \operatorname{argmax}_{IW \in \{IW' | POS(IW) \cap POS(UW) \neq \Phi\}} Sim(UW, IW) \quad (2)$$

Следует отметить, что наша конечная цель - улучшить качество перевода, но не все слова, найденные в словарном запасе с использованием формулы (2), могут гарантировать хороший перевод с контекстом неизвестного слова, поскольку им обычно не хватает соответствующей словарной записи в таблица фраз перевода. Таким образом, если найденное слово в словаре объединение контекста неизвестного слова соответствует записи в таблице фраз, это облегчит лексический выбор и переупорядочение слов в окружающих словах.

Практические результаты и программная реализация модели. Перевод и разделение сплошного текста на отдельные слова, исходный текст на английском языке и на русском языке. С помощью системы машинного перевода *googletrans* использованная в этом работе текст был переведен. Переведенный текст с помощью системы машинного перевода передается на NLTK(Tokenizer). Разделим текст на отдельные слова с помощью NLTK(Tokenizer).

После нормализации текст обрабатывается с помощью морфологического анализатора. Теггер анализирует разделенные слова, с помощью морфологического анализатора и пометить слова показано рисунке 2.

```
input sentence:
['Базалық', 'деңгейде', 'М', 'Т', 'бір', 'тілдегі', 'се
', ',', 'өйткені', 'аударма', 'тіліндегі', 'тұтас', 'ті

tagged sentence:
 1) Базалық        базалық_UNK
 2) деңгейде       деңгей_R_ZE де_C5
 3) М              m_UNK
 4) Т              t_UNK
 5) бір            бір_R_SN
 6) тілдегі        тіл_R_ZE дегі_LATT
 7) сөздерді       сөз_R_ZE дер_N1 ді_C4
 8) басқа          бас_R_ZE қа_C3
 9) тілдегі        тіл_R_ZE дегі_LATT
10) сөздерге       сөз_R_ZE дер_N1 ге_C3
11) механикалық    механикалық_UNK
12) ауыстыруды     ауыс_R_ET тыр_V4 у_ET_ETU ды_C4
13) жүзеге         жүзе_R_ZE ге_C3
14) асырады        асыр_R_ET а_T1 ды_P3
15) ,              ,_R_UTR
16) бірақ          бірақ_R_ZHL
17) тек            тек_R_US
18) жақсы          жақсы_R_SE
19) аударма        аударма_R_ZE
20) сирек          сирек_R_US
21) кездеседі      кездес_R_ET е_T1 ді_P7
```

Рисунок 2. Морфологический анализ текста и неизвестные слова

Теггер выбирает анализ, наиболее подходящий для данного контекста.

```
# создать экземпляр морфологического анализатора
analyzer = AnalyzerDD ()
# загрузить каталог модели, расположенный в каталоге морфологии
analyzer.load_model (os.path.join ('kaznlp', 'morphology', 'mdl'))
                                        wrd = 'ғалымы'
```

На рисунке 3 показан анализ заданного слова, для этого мы вызываем метод анализа со словом в качестве аргумента. метод возвращает кортеж (iscovered, alist), где iscovered - логическое значение, указывающее, было ли слово принято анализатором, alist - список результатов анализа, необработанные слова получат тег _UNK.

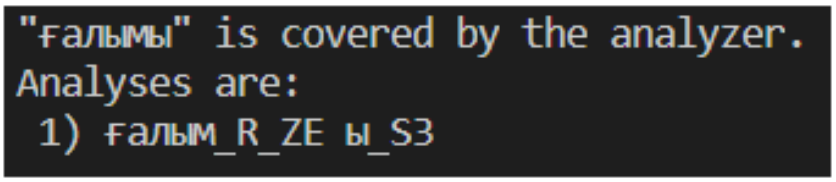

Рисунок 3. Анализ заданного слова

После анализа текста каждая слова получать. Каждый тэг имеет свое собственное обозначение, например, тег «R_ZE« - это существительное. Все данные исходного текста, переведенного текста, текст после морфологического анализа и неизвестные слова записывается в файлы. Работа написано на языке программирования Python и с помощью набора библиотек обработки текста. Последние результаты нахождения неизвестных слов с помощью морфологического анализатора поиск неизвестных слов для англо-казахского перевода составлен из 1000 предложений, взято более чем 18000 слов, в результате процент неизвестных слов составил 15%, а на втором поиске неизвестных слов было взято около 37000 слов в результате процент неизвестных слов составил 15,7%. Последние результаты нахождения неизвестных слов с помощью морфологического анализатора поиск неизвестных слов для русско-казахского перевода составлен из 1000 предложений, взято более чем 21693 слов, в результате процент неизвестных слов составил 14,02%, а на втором поиске неизвестных слов было взято около 43767 слов в результате процент неизвестных слов составил 14,12%.

Таблица 1 - **Результаты экспериментов по решению задачи неизвестных слов для англо-казахской и русско-казахской пары языков**

Языковая пара / Обработка	англо-казахская		русско-казахская	
Количество слов	18618	36999	21693	43767
Неизвестные слова	2800	5800	3042	6182
Процент неизвестных слов	15,04 %	15,67 %	14,02 %	14,12 %

Результаты нахождения неизвестных слов с помощью морфологического анализатора поиск неизвестных слов для англо-казахского перевода и для русско-казахского перевода, в поиске неизвестных слов было взято около 35000-45000 слов в результате процент нахождение неизвестных слов составил в среднем 14,5%. Таблице 1 результаты экспериментов по решению задачи неизвестных слов при постредактировании машинного перевода для англо-казахской и русско-казахской пары языков.

Заключение. В результате выполненных исследований были получены следующие результаты научно-технической деятельности:

- проведен обзор существующих методов нахождение неизвестных слов в постредактировании;
- разработана модель определения неизвестных слов в постредактировании машинного перевода;
- создан морфологический анализатор обработки слов для казахского языка;
- получены экспериментальные данные нахождения неизвестных слов в постредактировании машинного перевода для англо-казахской и русско-казахской языковой пары.

В итоге разработана модель и программная решения задачи неизвестных слов при постредактировании машинного перевода. Дальнейшее применение полученных результатов позволить улучшить процесс анализа текста на казахском языке, работу системы машинного перевода и постредактирования.

Благодарность. Исследование выполнено при поддержке Министерства образования и науки Республики Казахстан в рамках научного проекта AP08052421 «Исследование и разработка системы постредактирования казахского языка в машинном переводе».

ЛИТЕРАТУРА

[1] Matthias E., Stephan V., Alex W. Communicating Unknown Words in Machine Translation // 2014

[2] R. M. K. Sinha. Dealing with unknowns in machine translation // IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236), Tucson, AZ, USA, P. 940-944

[3] Zhang J., Zhai F., Zong Ch. Handling unknown words in statistical machine translation from a new perspective //Proceedings of the First CCF Conference Natural Language Processing and Chinese Computing P. 176–187 (2012)

[4] Turganbayeva A., Tukeyev U. The Solution of the Problem of Unknown Words Under Neural Machine Translation of the Kazakh Language // In: Intelligent Information and Database Systems 12th Asian Conference, P. 319–328 (2020)

[5] O. Makhambetov, A. Makazhanov, I. Sabyrgaliyev, Zh. Yessenbayev. Data-driven morphological analysis and disambiguation for Kazakh // InInternational Conference on Intelligent Text Processing and Computational Linguistics 2015, pp. 151-163.

[6] Luong M.T., Sutskever I., Le Q.V., Vinyals O., Zaremba W.: Addressing the rare word problem in neural machine translation // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 11–19 (2015)

[7] Hermann K.M. Teaching machines to read and comprehend // Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 1693–1701 (2015)

[8] Generating sequences with recurrent neural networks // https://arxiv.org/pdf/1308.0850.pdf.(дата обращения: 24.09.2020).

[9] Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1715–1725 (2016)

[10] Marton Y., Callison-Burch Ch., Resnik Ph.: Improved statistical machine translation using monolingually-derived paraphrases // Proceedings of the 2009 Conference on Empirical Methods in Natural Language, pp. 381–390 (2009)

[11] Zhang J., Zhai F., Zong Ch.: A substitution-translation-restoration framework for handling unknown words in statistical machine translation // J. Comput. Sci. Technol. 28(5), 907–918 (2013)

[12] Gulcehre C., Ahn S., Nallapati R., Zhou B., Bengio Y.: Pointing the unknown words // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 140–149 (2016)

[13] Li X., Zhang J., Zong C.: Towards zero unknown word in neural machine translation // Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2852–2858. AAAI Press (2016)

Д.Р. Рахимова, Н.М. Пазылхан*, А.А. Кульжанова, Ж.Г. Ален
әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
*e-mail: npazylhan@gmail.com

ПОСТРЕДАКЦИЯЛЫҚ МАШИНАЛЫҚ АУДАРМАДА БЕЛГІСІЗ СӨЗДЕРДІ АНЫҚТАУ МӘСЕЛЕСІНІҢ МОДЕЛІ МЕН БАҒДАРЛАМАЛЫҚ ШЕШІМІН ЖАСАУ

Андатпа. Машиналық аударма - компьютерлік бағдарламамен мәтіндерді бір тілден екінші тілге дәйекті түрде аудару технологиясы. Машиналық аударманың нәтижесінде әрдайым белгілі бір кемшіліктер бар, бұл мәселені постредакциялау арқылы шешуге болады. Пост-редакциялау - машиналық аудармадан кейін адамның мәтінді өңдеуі. Бүгінгі таңда көптеген лингвистикалық провайдерлер осы саланы белсенді дамытуда, редакторларды оқытудың әдістері мен пост-редакциялау әдістерін дамытуда. Мақалада постредакциялық машиналық аудармада белгісіз сөздерді табудың қолданыстағы әдістеріне шолу жасайды. Бұл жұмыста біз қазақ тіліне арналған постредакциялық машиналық аудармадағы белгісіз сөздерді анықтау мәселесін қарастырамыз. Редакциядан кейінгі машиналық аудармада белгісіз сөздерді табудың қолданыстағы әдістеріне талдау жүргізіледі. Ағылшын-қазақ және орыс-қазақ тілдеріне арналған постредакциялық машиналық аудармада белгісіз сөздерді табу моделі, практикалық нәтижелер және бағдарламалық қамтамасыз ету ұсынылған.

Негізгі сөздер: машиналық аударма, NLTK, морфологиялық талдау, белгісіз сөздер, постредакциялық машиналық аударма.

D.R. Rakhimova, N.M. Pazylkhan*, A.A. Kulzhanova, Zh.G. Alen
al-Farabi Kazakh national university, Almaty, Kazakhstan
*e-mail: npazylhan@gmail.com

DEVELOPMENT OF A MODEL AND SOFTWARE SOLUTION FOR THE PROBLEM OF DETERMINING UNKNOWN WORDS IN POST-EDITING MACHINE TRANSLATION

Abstract. Machine translation is the technology of consecutive translation of texts from one language to another by a computer program. As a result of machine translation, there are always certain disadvantages that can be solved by post-editing. Post-editing-human processing of text after machine translation. Today, many language providers are actively developing this field, developing methods of training editors and post-editing methods. The article provides an overview of existing methods for finding unknown words in post-editing machine translation. In this paper, we consider the problem of determining unknown words in post-editing machine translation for the Kazakh language. The analysis of existing methods for finding unknown words in post-editing machine translation is carried out. A model for the development of unknown words in post-editing machine translation for the English-Kazakh and Russian-Kazakh languages, practical results and software implementation are presented.

Keywords: machine translation, NLTK, morphological analysis, unknown words, machine translation post-editing.