МОДЕЛИРОВАНИЕ МОРФОЛОГИЧЕСКОГО АНАЛИЗА НА ОСНОВЕ ОКОНЧАНИЯ СЛОВА ДЛЯ УЗБЕКСКОГО ЯЗЫКА.

Шермухамедова Насиба Алмасовна

*Узбекистан. Бухара. Бухарский государственный университет,
филологический факультет, кафедра русского языка и литературы,
старший преподаватель*

Аннотация: *Узбекский язык, обладающий агглютинативным характером, образует слова путем сочетания аффиксов с корнями, используя флективные окончания для различных морфологических признаков. Это свойство создает большое количество комбинаций окончаний слов и значительно увеличивает словарный запас и объем данных проблемы разреженности статистических моделей. В этой статье рассматривается модель морфологического анализа, которая включает в себя выделение корней, лемматизацию и извлечение морфологической информации с учетом морфофонетических исключений. Суть модели заключается в разработке полного набора окончаний слов с соответствующей морфологической информацией и дополнительных наборов данных для морфологического анализа. Предложенная модель была протестирована с использованием специально разработанного набора тестов, включающего 5,3 тыс. слов. Точность на уровне слов превысила 91%, что было определено с помощью ручной проверки основы, леммы и исправлений морфологических признаков, проведенных экспертами-лингвистами. Созданный на основе предложенной методологии инструмент доступен в виде пакета Питон с открытым исходным кодом, а также веб-приложения, включающего общедоступный API.*

Ключевые слова: *узбекский язык, морфологический анализ, морфологическая сегментация, стемминг, лемматизация, связанная морфема, флективное окончание.*

MODEL ANALIZ MORFOLOZIK BAZE LO FIN MO POU LANG UZBEK.

Shermukhamedova Nasiba Almasovna

*Uzbekistan. Bukhara. Bukhara State University,
Faculty of Philology, Department of Russian Language and Literature,
Senior Lecturer*

Abstract. *Uzbek, an agglutinative language, forms words by combining affixes with roots, utilizing inflectional endings for various morphological features. This property makes a large number of combinations of word ending, and greatly increases the word-vocabulary size, and data sparseness problems for statistical models. This paper discusses a morphological analyzing model which includes stemming, lemmatizing and extraction of*

morphological information considering morpho - phonetic exceptions. A main point of the model involves developing a complete set of word-ending with assign morphological information, and additional datasets for morphological analysis. The proposed model was evaluated using a curated test set comprising 5.3K words. It achieved a word-level accuracy over 91%, as determined through manual verification of stem, lemma, and morphological feature corrections conducted by linguistic experts. The created tool based on the proposed methodology is available as an open-source Python package, as well as a web-based application including a public API.

Keywords: *uzbek language, morphological analyzing, morphological segmentation, stemming, lemmatizing, bound morpheme, inflectional ending.*

O'ZBEK TILI UCHUN SO'ZNING OXIRI ASOSIDA MORFOLOGIK TAHLILNI MODELLASHTIRISH.

Shermuhamedova Nasiba Almasovna

*O'zbekiston. Buxoro. Buxoro davlat universiteti,
Filologiya fakulteti, rus tili va adabiyoti kafedrası,
katta o'qituvchi*

Annotatsiya: *Agglyutinatív xususiyatga ega bo'lgan o'zbek tilida affikslarni o'zak bilan bog'lash, turli morfologik belgilarga ko'ra qo'shimchalar qo'llash orqali so'z hosil qiladi. Bu xususiyat juda ko'p sonli so'z birikmalarini yaratadi va statistik modellarning siyraklik muammosining so'z boyligi va ma'lumotlar hajmini sezilarli darajada oshiradi. Ushbu maqolada morfofonetik istisnolarni hisobga olgan holda ildizni ajratib olish, lemmatizatsiya va morfologik ma'lumotni ajratib olishni o'z ichiga olgan morfologik tahlil modeli muhokama qilinadi. Modelning mohiyati mos keladigan morfologik ma'lumotlar va morfologik tahlil uchun qo'shimcha ma'lumotlar to'plami bilan so'zlarning to'liq to'plamini ishlab chiqishdir. Taklif etilayotgan model maxsus ishlab chiqilgan testlar to'plami, jumladan 5,3 ming so'z yordamida sinovdan o'tkazildi. So'z darajasidagi aniqlik 91% dan oshdi, bu tilshunoslar tomonidan qo'lda o'zak, lemma va morfologik xususiyatlarni tuzatish orqali aniqlandi. Taklif etilgan metodologiya asosida yaratilgan vosita ochiq manba Python paketi, shuningdek, umumiy API ni o'z ichiga olgan veb-ilova sifatida mavjud.*

Kalit so'zlar: *o'zbek tili, morfologik tahlil, morfologik segmentatsiya, o'zak, lemmatizatsiya, bog'langan morfema, flektiv son.*

Вступление. Компьютерная лингвистика интегрирует моделирование естественного языка человека с помощью основанного на правилах подходы, основанные на статистических моделях, моделях машинного и глубокого обучения. Тот факт, что в некоторых языках широко используются суффиксы и префиксы для передачи грамматического значения (например, согласование подлежащего и глагола), представляет собой проблему для большинства современных

языковых технологий. Суффиксы и префиксы в таких языках в более общем смысле можно назвать морфемами, которые определяются как значимые составные части слов. Правила, которые языки используют для сочетания морфем, а также сами используемые ими морфемы называются морфологией языка. Языки, которые широко используют морфемы для построения слов, считаются морфологически богатыми. Узбекский язык, как и другие тюркские языки, имеет богатую морфологию и множество исключений в морфологической структуре слов. Морфология фокусируется на том, как части слова, такие как основы, префиксы и суффиксы, организованы или изменены для передачи различных значений. Существует две основные категории морфологии: словоизменяющая морфология и словообразовательная морфология. Каждая из этих двух категорий имеет свое значение в различных аспектах НЛП. Словоизменяющие морфемы, как правило, суффиксы, добавляются до конца слова, чтобы передать грамматические нюансы. Эти словоизменяющие окончания служат для изменения значения слова, сохраняя при этом его фундаментальное значение. В этом проекте мы исследуем модель морфологического анализа, основанную на полном наборе окончаний слов, которые образовались только из словоизменяющих морфем. Модель морфологического анализа включает в себя методы стемминга, лемматизации и морфологической сегментации наряду с извлечением морфологической информации. Морфологический анализ узбекского языка предполагает разбор сложных структур слов, включающих префиксы, суффиксы и корневые формы. Высокая флективность языка требует тщательного подхода к выделению признаков и пониманию контекстуальной семантики. Модели нейронных сетей, обладающие способностью улавливать сложные паттерны и контекстуальные свойства, обладают потенциалом для улучшения работы в этом сложном лингвистическом ландшафте.

Обзор литературы. Работа [1] была посвящена разработке морфологических анализаторов, которые разделяют слова на последовательности морфем или слогов в уйгурском языке. Процесс сегментации морфем является центральным в создании всеобъемлющего корпуса уйгурского языка. В то время как контролируемый подход в сочетании с правилами и алгоритмами статистического обучения эффективно справляется с сегментацией морфем, значение лингвистических морфем в расширении охвата слов и сокращении размера словарного запаса. Кроме того, эти морфологические анализаторы предназначены для обработки как стандартных, так и поверхностных форм, учитывающих фонетические изменения и сложные морфологические изменения.

Хранение всех поверхностных слов в словаре для морфологически богатого тюркского языка уже давно является сложной задачей и накладывает ограничения на широкое применение нейронного машинного перевода (NMT). В работах [2], [3] устраняется это ограничение путем внедрения инновационного подхода к морфологической сегментации тюркских языков, основанного на концепции полного набора окончаний (CSE). Основанный на CSE метод сегментации, как показано на

примере казахского, кыргызского и узбекского языков, эффективно уменьшает объем словарного запаса в исходных текстах. Вычислительные эксперименты NMT, в частности, с акцентом на казахском языке, показали убедительные результаты. По сравнению с сегментацией, основанной на парном байтовом кодировании (BPE), подход, основанный на CSE based, повышает балл двуязычного дублера для казахско–английских и англо–казахских пар. В контексте морфологического анализа узбекского языка было предложено несколько моделей с использованием конечных автоматов (FSM). В [4] авторы используют предопределенные грамматические правила для извлечения морфологической информации из заданных слов с помощью автоматов. Однако важно отметить, что эта методология использует словарь словоформ в качестве фундаментальной базы данных, в отличие от использования грамматических, морфотаксических и морфологических правил. В работе [5], которая является одной из работ по морфологическому анализу слов узбекского языка на основе модели конечного автомата (FSM), исследуется задача определения корня с извлечением соответствующей морфологической информации. Тем временем для узбекского языка были разработаны морфотаксические правила [6] и модели морфологического анализа [7], [8]. Считается, что узбекские слова имеют довольно сложную структуру, и существует множество исключительных случаев. Узбекский язык представляет собой сложную задачу из-за наличия омонимии и синонимии аффиксов, гармонии гласных, что может привести к ошибкам в морфологической сегментации. Это также может привести к снижению точности моделей морфологического анализа, основанных на FSM. Методология. В этом разделе мы представляем подробное описание узбекской модели морфологического анализа. Используемая модель дополняет набор данных об окончаниях слов и дополнительных наборах данных и включает в себя методы определения корней, лемматизации и морфологического анализа на основе правил.

Начальный этап предлагаемой методологии включает в себя разработку полного набора языковых данных по окончаниям (CSE). Морфемная часть слова будет разделена на две части: словообразовательную и словоизменительную морфологию. Словоизменительная морфология включает в себя последовательности словосочетаний, лексические и синтаксические суффиксы. Словообразовательные морфемы изменяют часть речи слова. В этой работе мы сохраняем словоизменительную часть слова. При формировании набора данных мы собрали все возможные окончания слов; была предоставлена морфологическая информация, связанная с окончаниями каждого слова. Эта морфологическая информация включает различные лингвистические атрибуты, такие как часть речи (POS), время, принадлежность, связка, форма единственного или множественного числа, форма вопроса, падеж и многое другое. Набор данных окончаний слов создается экспертами-лингвистами для обеспечения точности морфологической информации. На основе анализа морфологической структуры слов была разработана структура

базы данных для хранения всей необходимой информации для модели морфологического анализа.

Из-за гармонии гласных и правил аффиксации суффикс в узбекском языке может иметь несколько алломорфов. Итак, в наборе данных окончаний слов отмечены конкретные обозначения для учета вариаций и алломорфов. Для описания алломорфных суффиксов мы используем следующие обозначения из предыдущей работы [5]:

G: {g,k,q}; K: {g,k}; Q: {g,g',k,q}; Y: {a,y}; T: {t,d}; () : буква, заключенная в круглые скобки, может быть опущена.

Для процесса ввода мы разработали веб-приложение, в котором каждому эксперту-лингвисту присваиваются уникальные имя пользователя и пароль, с помощью которых он может получить доступ к веб-сайту и ввести новое слово, заканчивающееся соответствующей морфологической информацией. Значительная часть словоизменительных окончаний правильно соответствует глаголу и существительному как грамматической функции слова. В наборе данных 1205 и 150 записей являются глаголами и существительными соответственно. Кроме того, набор данных содержит 22 записи числительных, 10 записей прилагательных, 20 записей местоимений и 10 записей наречий. Кроме того, мы сформировали пять наборов данных для повышения точности производительности методологии. Исключительные корни (доллар, ташкари, ...), список корней, для которых в нем есть какой-либо дополнительный конец. Неаффиксированные основы (va, lekin, ya na, ...), список основ, к которым никогда не было никаких прикреплений.

Числовые основы (bir, ikki, ellik, ...).

Основы короткой длины (длина $\leq$ 2) (u, bu, ot, un, ...).

Исключительная лемма (bitta, bir, ta / single, singil, i).

На следующем этапе методологии для каждого элемента будут сгенерированы алломорфы из столбца affix в наборе данных CSE. Затем в процесс моделирования будут включены спецификации алфавита, которые используются для проверки морфологических правил. После этого вводимые слова будут перейти к процессу очистки, который включает в себя преобразование всех символов в нижний регистр и замену определенных специальных символов по мере необходимости. Это гарантирует, что введенные слова будут надлежащим образом подготовлены для последующей обработки и проверки правил.

Прежде чем определить окончание введенного слова, мы проверяем предопределенные основы в дополнительных созданных наборах данных, что гарантирует быструю работу модели и низкий уровень ошибок. Извлекается окончание слова, что предоставляет дополнительный уровень морфемной информации. Для повышения точности морфологического анализа при извлечении окончаний слов применяются правила аффиксации. Затем извлеченное окончание слова сравнивается с набором данных CSE для выявления совпадений. При

обнаружении совпадения результирующему набору присваивается соответствующая морфологическая информация. Когда прогнозируемая длина стебля составляет менее двух знаков, выполняется дополнительный этап проверки. Для подтверждения правильности определения стебля используется набор данных о короткой длине стебля. Этот процесс проверки обеспечивает точность определения коротких стеблей, что повышает общую надежность морфологического анализа. Результаты этих этапов обработки возвращают список, включающий основу, лемму и морфологическую информацию для данного слова и POS-тега. Результаты и обсуждения. В качестве цели оценки предложенная модель была проанализирована с использованием данных построенного корпуса для метода морфологической сегментации. Для тестирования был использован построенный корпус, состоящий из 40 документов, полученных с новостной платформы (daryo.uz), которые были разделены на четыре группы по десять документов в каждой. Этот корпус содержит в общей сложности 11 952 слова, из которых 5 288 уникальных составляют 44,24% от общего числа слов. Слова были аннотированы экспертом-лингвистом с правильной сегментацией. В процессе оценки время работы модели составляет около 1000 слов в секунду. Выходные данные модели сравниваются для каждого слова с фактическими ожидаемыми выходными данными. Выходные данные модели были разделены на пять различных вариантов, один из которых указывал на правильные прогнозы, в то время как остальные четыре указывали на различные типы ошибок, обнаруженных во время оценки. В таблице 1 результаты оценки модели на основе собранного корпуса демонстрируют разбивку результатов по категориям в случаях. В таблице поверхностное слово определено как образец китобимдан (из моей книги), оно разделено на три морфемы: китоб [основа], -(i)м [принадлежность к первому лицу] и -дан [аблятивный падеж].

Примерный пример кейсов %

1. Правильное предсказание китобимни > китоб 4821 91,2%

2. Ни один аффикс не был вычеркнут, хотя аффикс(ы) существует китобимни > китобимни 24 0,5%

3. Аффикс(ы) был удален из

основы, в то время как основа имеет аффикс(ы) китобимни > кито 131 2,5%

4. Частично полосатые аффиксы китобимни > китобим 158 3,0%

Большинство ошибок возникает из-за наличия коротких окончаний слов в основах (3-й, 5-й падежи), и все варианты еще не охвачены в наборе данных CSE (4-й случай). Инструмент Питон, созданный для этой работы, находится в открытом доступе, а также может быть легко установлен с помощью следующей команды, популярной в сообществе Python: `pip install морфологический_анализатор`.

Существует также общедоступная система API 2 для интеграции модели в другое программное обеспечение, и более подробную информацию о ней можно найти в репозитории проекта на гитхаб 3.

Вывод. В этом исследовании проводится всесторонний морфологический анализ узбекского языка, охватывающий различные задачи, связанные со структурой слов и морфологическими свойствами. Методология включает в себя разработку набора данных о словоизменительных окончаниях, который лингвистические эксперты дополняют морфологической информацией. помощью существительных и глаголов, которые играют определенную роль в изменении грамматических свойств слов. Процесс оценки включает в себя определение основы, лемматизацию и морфологический анализ, что приводит к впечатляющему показателю точности – более 91% для проверки основы и леммы. Кроме того, мы представили код на Питон, веб-инструмент и API, созданные для использования предложенной модели.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА:

1. М. Аблимит, Т. Кавахара, А. Паттар и А. Хамдулла, “Уйгурский язык, основанный на ствольных аффиксах” Morphological Analyzer, ”Международный журнал коммуникации будущего поколения и Сетевое взаимодействие, том 9, № 2, 2016, doi: 10.14257/ijfgcn.2016.9.2.07.

2. У. Тукеев, А. Карибаева и З. Х. Жуманов, “Метод морфологической сегментации Нейронный машинный перевод на тюркский язык”, Cogent Eng, том 7, № 1, стр. 1856500, 2020, doi: 10.1080/23311916.2020.1856500.

3. Шермухамедова Н. А. Графика и транскрипция заимствованных имен // Barqaror Taraqqiyot va Rivojlanish Tamoyillari. – 2023. – Т. 1. – №. 1. – С. 34-36.

4. Шермухамедова Н. А. Лингвистические проблемы перевода: текст как объект переводческой деятельности // POLISH SCIENCE JOURNAL. – 2020. – С. 335.

5. Шермухамедова Н. А. Ходжи на Насреддина как источник духовно-нравственно воспитания молодёжи // Тошкент-2031- С.31.

6. Мукимова Г. А. оптимальные пути обучения иностранному языку: Янги Узбекистон тараккиотида таджикотлар урни ва ривожланиш омиллари // Янги Узбекистон тараккиотида таджикотларни о’рни ва ривойланиш омиллари. – 2024. – Т. 4. – №. 1. – С. 159-163.

7. Шермухамедова Н.А. словообразование аспект компьютерного слена и интернет-дискурса. 8. Туйлиева Л. А. Методы. Методы исследования глаголов состояния: // Образование и инновационные исследования международный научно-методический журнал. – 2023. – №. 3. – С. 65-73.

9. Истамовна Д. С. Категории оппозиции и контраста как объекты лингвистического анализа // EPRA International Journal of Multidisciplinary Research (IJMR). – 2023. – Т. 9. – №. 11. – С. 407-409.

10. Аvezова Д. литературные псевдонимы на примере русских и узбекских писателей // Наука и инновации. – 2024. – Т. 3. – №. С3. – С. 90-95.