ГИБРИДНАЯ ФОРМАЛЬНО-НЕЙРОННАЯ МОДЕЛЬ СИНТАКСИЧЕСКОГО РАЗБОРА ПРЕДЛОЖЕНИЙ ОФИЦИАЛЬНО-ДЕЛОВОГО СТИЛЯ УЗБЕКСКОГО ЯЗЫКА: АРХИТЕКТУРА, ФОРМАЛИЗАЦИЯ И ЭКСПЕРИМЕНТАЛЬНАЯ МЕТОДИКА

Сайдаматов Махмуджон

доцент кафедра искусственного интеллекта и естественных наук, Университет геологических

Жураева Нилуфар Вайитовна

доцент кафедра искусственного интеллекта и естественных наук, Университет геологических наук

Жураева Умида Вохидовна

преподаватель кафедры экономика и туризма Университета Афрасиаб.

A HYBRID FORMAL-NEURAL MODEL FOR SYNTACTIC PARSING OF UZBEK OFFICIAL-ADMINISTRATIVE STYLE SENTENCES: ARCHITECTURE, FORMALIZATION, AND EXPERIMENTAL METHODOLOGY

Makhmudjon Saydamatov

Associate Professor, Department of Artificial Intelligence and Natural Sciences, University of Geological Sciences

Nilufar Vayitovna Juraeva

Associate Professor, Department of Artificial Intelligence and Natural Sciences, University of Geological Sciences

Umida Vohidovna Juraeva

Lecturer, Department of Economics and Tourism, Afrosiyob University

Аннотация

В Узбекистане цифровизация госуправления приводит к значительному увеличению официальных текстов, которые нужно быстро и автоматически обрабатывать - при этом важно, чтобы результат был понятным и объяснимым. Классические формальные грамматики дают строгий и проверяемый разбор, но плохо справляются с гибкостью языка и омонимией. Нейросети, наоборот, устойчивы к этим проблемам, но не гарантируют корректный синтаксис и остаются «чёрным ящиком».

Чтобы объединить сильные стороны обоих подходов, предлагается гибридную модель: нейронный модуль генерирует варианты разбора, формальная грамматика проверяет их на правильность, а ранжирующий блок выбирает лучший и объясняет, почему именно он. Такая система особенно полезна для обработки нормативно-правовых документов и создания объяснимых инструментов NLP.

Ключевые слова: *узбекский язык; официально-деловой стиль; синтаксический анализ; формальная грамматика; аксиоматическая модель; нейросимвольный ИИ; NLP.*

Abstract

The digitalization of public administration in Uzbekistan has led to a significant increase in the volume of official texts that require rapid and automated processing. At the same time, it is essential that the results of such processing are understandable and explainable. Classical formal grammars provide a rigorous and verifiable syntactic analysis but often struggle with linguistic flexibility and ambiguity. Neural networks, in contrast, are more robust in addressing these challenges; however, they do not always guarantee syntactic correctness and often remain a “black box.”

To combine the strengths of both approaches, this study proposes a hybrid model in which a neural module generates possible syntactic parses, a formal grammar verifies their correctness, and a ranking component selects the most appropriate parse while explaining the reasons for its selection. Such a system is particularly useful for processing regulatory and legal documents and developing explainable natural language processing (NLP) tools.

Keywords: Uzbek language; official-administrative style; syntactic analysis; formal grammar; axiomatic model; neuro-symbolic artificial intelligence; natural language processing (NLP).

В Узбекистане госуправление постепенно переходит на электронный документооборот. Это значит, что растут требования к автоматической обработке текстов на узбекском языке. Официально-деловой стиль в этом плане удобен: он формализован и содержит устойчивые конструкции, что позволяет сочетать формальные и статистические методы. Но есть и сложность - агглютинативная морфология создаёт множество омонимов и мешает напрямую переносить решения с других языков.

Главная проблема - противоречие двух подходов. Формальные модели дают строгий и объяснимый синтаксический разбор, но плохо справляются с вариативностью текста. Нейросети, наоборот, обеспечивают высокую точность, но не гарантируют грамматическую правильность и остаются «чёрным ящиком». В официально-деловом стиле, где цена ошибки особенно высока, это становится серьёзным ограничением.

Научная проблема состоит в объединении этих двух свойств в одном анализаторе. В предлагаемой архитектуре нейронная модель не заменяет формальную грамматику. Она генерирует кандидатов. Формальная система проверяет их допустимость. Затем ранжирующий блок выбирает наиболее вероятную структуру среди прошедших проверку. Для достижения поставленной цели основные задачи заключаются в следующем: аксиоматическая грамматика рассматривается как формальный модуль проверки, при этом 12 правил вывода и множество порождаемых ими конструкций образуют контролируемое пространство синтаксических классов; далее разрабатывается нейронный модуль генерации кандидатов, задаётся математическая функция совместной оценки, после чего проводится сравнение формальной, нейронной и гибридной моделей в одинаковых условиях, а завершающим этапом выступает анализ ошибок по типам синтаксических конструкций

Формальные грамматики дают строгую и проверяемую процедуру разбора. Их ценность для официально-делового текста в том, что результат можно проверить независимо от статистических моделей. В этой работе формальная система строится вокруг аксиомы правильного порядка компонентов предложения: «подлежащее + дополнение + обстоятельство + сказуемое». На основе этой аксиомы вводятся 12 правил вывода, которые позволяют описывать разные типы предложений - от простых и определительных до вопросительных, отрицательных, сложносочинённых и сложноподчинённых, включая конструкции с прямой и косвенной речью.

Нейронные методы работают иначе: они создают контекстное представление, учитывающее не только отдельное слово, но и его окружение. Это особенно важно для узбекского языка, где агглютинативная морфология порождает омонимию и неоднозначные формы.

Отсюда возникает исследовательский разрыв: формальная система и нейронная система решают разные части задачи. Предлагаемый подход объединяет их на уровне принятия решения - нейросеть генерирует кандидатов, формальная грамматика проверяет их допустимость, а итоговый выбор делается совместно, а не как результат двух независимых моделей.

Формальная модель представляется как теория U. Её базовая аксиома задаёт правильный порядок компонентов:

1. **Формальная модель**

A : S + O + Adv + P (*Rahbariyat topshiriqni belgilangan muddatda bajardi*.**)**

где S - подлежащее, O - дополнение, Adv - обстоятельственный компонент, P - сказуемое.

Система правил вывода имеет вид: $R = \{r_1, r_2, \ldots, r_{12}\}$и имеет 12 порождаемых типов конструкций: PV.1 - нераспространённое предложение; PV.2 - предложение с определением; PV.3 - обще-вопросительное; PV.4 - специальный вопрос; PV.5 - негативное; PV.6 - сложносочинённое; PV.7 - прямая речь с конструкцией A+P; PV.8 - прямая речь A+P+A; PV.9 - прямая речь P+A+P; PV.10 - прямая речь P+A; PV.11 - косвенная речь; PV.12 - сложноподчинённое предложение.

Для формальной проверки кандидата y_i вводится индикатор:

$$V(y_i) = \begin{cases} \mathbf{1}, \text{если } \mathbf{y_i} \text{ выводим по правилам } \mathbf{R} = \{\mathbf{r_1}, \mathbf{r_2} \ldots \mathbf{r_{12}}\} \\ \mathbf{0}, \text{иначе} \end{cases}$$

При строгом режиме анализатор принимает только структуры, для которых $V(y_i)=1$. Преимущество такого решения состоит в том, что система может сохранить цепочку правил, которая привела к принятому результату.

2. Нейронная модель

Нейронный компонент рассматривается как генератор синтаксических гипотез. На вход подаётся токенизированное предложение $X=(x_1,\ldots,x_n)$ ($X = (x_1, x_2, x_3, x_4)$ = ("Talaba", "kutubxonada", "kitob", "o'qiydi")). Контекстное представление вычисляется функцией $H=f_\theta(X)$ ($H = \{h_1, h_2, \ldots, h_k\}$ - множество гипотез, где каждая h_i может быть, например, вариантом дерева зависимостей: h_1 - ("Talaba" = подлежащее, "kutubxonada" = обстоятельство места, "kitob" = дополнение, "o'qiydi" = сказуемое). На основе H модель формирует k кандидатов и соответствующие вероятности: $Y = \{y_1,\ldots,y_k\}$, $P(yi|X)$. Эта формула описывает следующий шаг: на основе множества гипотез H нейронная модель формирует k конкретных кандидатов-разборов $Y = \{y_1, \ldots, y_k\}$ и для каждого из них вычисляет вероятность $P(y_i|X)$ - насколько правдоподобен именно этот вариант синтаксического разбора при данном входном предложении X. Каждому y_i ставится в соответствие число $P(y_i|X) \in [0,1]$, причём сумма всех вероятностей по k кандидатам равна единице ($\Sigma\ P(y_i|X) = 1$), поскольку модель распределяет вероятностную массу между конкурирующими интерпретациями. Применительно к узбекскому примеру X = ("Talaba", "kutubxonada", "kitob", "o'qiydi"): Кандидаты $Y = \{y_1, y_2, y_3\}$:

- y_1: подлежащее–обстоятельство–дополнение–сказуемое (Talaba–kutubxonada–kitob–o'qiydi), с вероятностью $P(y_1|X) = 0.91$
- y_2: альтернативная интерпретация, где "kutubxonada" трактуется как часть составного дополнения, с вероятностью $P(y_2|X) = 0.05$
- y_3: маловероятный разбор с иной группировкой словосочетаний, с вероятностью $P(y_3|X) = 0.02$

В отличие от чисто нейронного подхода, самый вероятный кандидат не принимается автоматически. Каждый кандидат передаётся в формальный модуль. Для нейронной модели множество синтаксических классов задаётся

формальной системой. Это делает эксперимент сопоставимым: нейронная и гибридная модели решают одну и ту же задачу структурного выбора, а различаются наличием формальной проверки.

3. Гибридная архитектура

Общая последовательность обработки имеет вид:

X → Morphology → Neural Proposal → Formal Verification → Neural Ranking → Explainable Result

"Talaba" (сущ., им. п., ед. ч.), "kutubxonada" (сущ. + локативный аффикс -da), "kitob" (сущ., им./вин. п.), "o'qiydi" (глаг., наст.-буд. вр., 3 л. ед. ч.).

Сначала выполняется предварительная обработка. Затем нейронная модель предлагает несколько структур. Формальный модуль проверяет каждый кандидат по 12 правилам. Неверифицируемые структуры удаляются в строгом режиме. Оставшиеся кандидаты ранжируются.

Нейронная модель предлагает кандидатов Y = {y_1, y_2, y_3}:

- y_1: Talaba(подлежащее) - kutubxonada(обстоятельство места) - kitob(дополнение) - o'qiydi(сказуемое), $P(y_1|X) = 0.91$
- y_2: Talaba(подлежащее) - kutubxonada+kitob(составное дополнение) - o'qiydi(сказуемое), $P(y_2|X) = 0.05$
- y_3: Talaba(подлежащее) - kitob(подлежащее второго уровня, ошибочно) - kutubxonada - o'qiydi(сказуемое), $P(y_3|X) = 0.02$

Множество допустимых кандидатов определяется как:

$Y' = \{ y_i \in Y : V(y_i)=1 \}$

Формальный модуль проверяет каждого кандидата по 12 правилам вывода, присваивая бинарную верификацию $V(y_i)$:

- $V(y_1) = 1$ - структура полностью выводима: подлежащее в им. п. согласуется со сказуемым по лицу/числу, обстоятельство места оформлено локативным аффиксом, дополнение стоит перед глаголом (SOV-порядок для узбекского языка) - все 12 правил удовлетворены.
- $V(y_2) = 0$ - конструкция "kutubxonada+kitob" как единое составное дополнение не выводится ни одним из 12 правил (нет правила, объединяющего локатив и прямое дополнение в одну группу без союза).
- $V(y_3) = 0$ - нарушено правило единственности подлежащего: две именные группы претендуют на роль подлежащего, что противоречит аксиомам грамматики.

Итоговый выбор можно задать:

$$y* = arg \max_{y_i \in Y'} \text{Score}(y_i)$$

Строгий режим - удаление неверифицируемых кандидатов:

$Y' = \{ y_i \in Y : V(y_i) = 1 \} = \{ y_1 \}$

Поскольку y_2 и y_3 не прошли формальную проверку, они исключаются, и в множестве допустимых кандидатов остаётся только y_1.

Ранжирование и итоговый выбор:

$$y* = arg \max_{y_i \in Y'} \text{Score}(y_i) = y_1$$

Поскольку Y' содержит только один элемент, y_1 автоматически становится итоговым выбором (в случае, если бы в Y' осталось несколько кандидатов, $\text{Score}(y_i)$ мог бы учитывать, например, комбинацию $P(y_i|X)$ и структурных признаков).

Для расширенного режима допускается мягкая формальная оценка $C(y_i) \in [0,1]$. Это позволяет работать с предложениями, которые частично выходят за пределы контролируемого фрагмента.

Расширенный режим с мягкой формальной оценкой $C(y_i) \in [0,1]$.

Здесь вместо жёсткого $V(y_i) \in \{0,1\}$ формальный модуль выдаёт степень соответствия правилам, что полезно для предложений, частично выходящих за пределы контролируемого фрагмента (например, с эллипсисом или разговорным порядком слов):

- $C(y_1) = 1.00$ (полностью соответствует всем 12 правилам)
- $C(y_2) = 0.40$ (частично нарушает правило оформления дополнения, но синтаксически распознаваемо)
- $C(y_3) = 0.15$ (грубое нарушение правила единственности подлежащего, почти не выводимо)

4. Составная функция оценки

Итоговая оценка кандидата определяется как выпуклая комбинация трёх источников информации:

$$Score(T_i) = \alpha * S_{neural}(T_i) + \beta * S_{formal}(T_i) + \gamma * S_{semantic}(T_i)$$

$\alpha + \beta + \gamma = 1, \quad \alpha,\beta,\gamma \geq 0$

где T_i - рассматриваемая синтаксическая конструкция; $S_{neural}(T_i)$ - оценка правдоподобия конструкции нейронной моделью; $S_{formal}(T_i)$ - оценка степени соответствия конструкции формальным правилам вывода; $S_{semantic}(T_i)$ - оценка согласованности семантических ролей внутри конструкции; Коэффициент α отражает степень доверия к статистическому правдоподобию, β - вес требования формальной корректности, γ - вклад семантической согласованности членов предложения. В строгом режиме формальная проверка является не только частью оценки, но и обязательным ограничением: кандидаты с $V(y_i)=0$ не участвуют в окончательном выборе.

В этом режиме итоговый $\text{Score}(y_i)$ в упрощённом виде может, например, комбинировать только нейронную и формальную оценки: $\text{Score}(y_i) = \alpha \cdot P(y_i|X) + (1-\alpha) \cdot C(y_i)$, где α - весовой коэффициент гибридной модели. Это частный случай общей составной функции оценки, вводимой в разделе 4 (при $\gamma=0$ и $\beta=1-\alpha$), полезный, когда семантическая компонента не учитывается отдельно. Выбор y^* делается по максимуму этой совместной функции уже без жёсткого отсечения кандидатов.

Обучение нейронного компонента производится минимизацией эмпирической функции потерь:

$$L(\theta) = \frac{1}{N} \cdot \sum_{i=1}^{N} \ell(y_i, \hat{y}_i)$$

L(θ) - средняя ошибка модели при параметрах θ;
N - количество обучающих примеров;
i - номер примера;
yi - истинный результат (эталонная разметка);
ŷ - результат, предсказанный моделью;
$\ell(y_i, \hat{y}_i)$ - функция ошибки для одного примера;
$\sum_{i=1}^{N} l(y_i, \hat{y}_i)$ - суммирование ошибок всех примеров;
$\frac{1}{N}$ - вычисление **средней ошибки**.

Для полноценной проверки необходим специализированный корпус официально-деловых текстов узбекского языка. В исходной работе в качестве источников указаны нормативно-правовые и административные документы, письма, постановления, распоряжения, договоры и государственные бланки.

Разметку целесообразно выполнять на трёх уровнях: морфология, синтаксис и тип конструкции. Предложения одного документа не должны одновременно попадать в разные выборки, чтобы избежать утечки информации.

5. Расчёт качества по типам конструкций

Каждый тип конструкции рассматривается отдельно. Для класса C_j рассчитывается точность:

$$\text{Accuracy}(C_j) = \frac{\text{Correct}(C_j)}{N(C_j)}$$

Для сравнения моделей вводятся три значения: $A_F(C_j)$- точность формальной (правило-ориентированной) модели
, $A_N(C_j)$- точность нейронной модели
и $A_H(C_j)$- точность гибридной модели (нейронная + формальная проверка)

Прирост гибридной модели определяется как:

$\Delta_{H-N}(C_j) = A_H(C_j) - A_N(C_j)$- насколько формальная проверка улучшила работу нейронной модели на конструкции Cj.

$\Delta_{H-F}(C_j) = A_H(C_j) - A_F(C_j)$- насколько гибридная модель лучше чисто формальной на этой же конструкции.

Здесь важно понимать: точность $A(C_j)$ - это метрика, усредняемая по множеству предложений одного типа конструкции, а не по одному предложению. Поэтому наше предложение "Talaba kutubxonada kitob o'qiydi" может выступать лишь как один представитель (один тестовый пример) внутри какого-то класса C_j, а сама точность вычисляется на всей группе таких предложений.

Обычное сравнение по F1 даёт только общую оценку «модель А лучше модели В». А эта схема позволяет ответить на более интересный вопрос: **Для каких**

именно конструкций добавление формальной проверки реально снижает количество ошибок нейронного анализатора? То есть можно увидеть, где нейронная модель «плавает», а формальные правила её успешно подстраховывают.

a. **Определение класса конструкции.** Предложение "Talaba kutubxonada kitob o'qiydi" можно отнести, например, к классу C_j = "простое предложение с обстоятельством места, выраженным местным падежом" (подлежащее + обстоятельство места на -da + дополнение + сказуемое). Пусть в тестовой выборке есть, скажем, 50 таких предложений этого типа, и наше - одно из них.
b. **Расчёт точности $A(C_j)$ для каждой модели.** Точность класса - это доля правильно разобранных предложений этого типа из общего числа предложений данного класса:

$A(C_j)$ = (число верно разобранных предложений класса C_j) / (общее число предложений класса C_j)

Для нашего предложения, допустим, все три модели дают следующий результат на нём: формальная модель, применяя 12 правил вывода, безошибочно строит y_1 (Talaba=подлежащее, kutubxonada=обстоятельство места, kitob=дополнение, o'qiydi=сказуемое) - верно. Нейронная модель предсказывает $\hat{y} = y_2$ (ошибочно объединяет kutubxonada и kitob в составное дополнение) - неверно. Гибридная модель берёт нейронного кандидата y_2, отсекает его формальной проверкой ($V(y_2)=0$) и возвращает y_1 - верно.

Если усреднить по всем 50 предложениям класса C_j (наше - лишь один пример из них), могли бы получиться, например, такие значения:

c. $A_F(C_j) = 0.88$ (формальная модель верно разбирает 44 из 50 предложений - ошибается там, где встречаются конструкции, не покрытые 12 правилами)
d. $A_N(C_j) = 0.74$ (нейронная модель верно разбирает 37 из 50 - путает локатив с дополнением, как в нашем примере с y_2)
e. $A_H(C_j) = 0.94$ (гибридная модель верно разбирает 47 из 50 - формальная проверка отсеивает часть ошибок нейронной сети, включая случай с нашим предложением)
f. **Прирост гибридной модели относительно нейронной:**
   $\Delta_{H-N}(C_j) = A_H(C_j) - A_N(C_j) = 0.94 - 0.74 = +0.20$
   Это значит, что для конструкций типа "подлежащее + локативное обстоятельство + дополнение + сказуемое" формальная проверка исправила 20 процентных пунктов ошибок нейронной модели - в том числе и тот самый случай, когда нейросеть ошибочно предпочла y_2, а формальный модуль вернул разбор к правильному y_1.
g. **Прирост гибридной модели относительно формальной:**

$\Delta_{H-F}(C_j) = A_H(C_j) - A_F(C_j) = 0.94 - 0.88 =$ **+0.06**

Это означает, что гибридная модель превзошла чисто формальную модель на 6 процентных пунктов - вероятно, за счёт случаев, где формальные правила

допускали несколько верифицируемых кандидатов ($V(y_i)=1$ для нескольких y_i), и именно нейронная вероятность $P(y_i|X)$ помогла выбрать наиболее вероятный из них, тогда как чисто формальный модуль без ранжирования мог выбрать не оптимальный вариант. Как и во всех предыдущих примерах, конкретные числа (0.88, 0.74, 0.94) - иллюстративные; реальные значения $A(C_j)$, $\Delta_{H-N}(C_j)$ и $\Delta_{H-F}(C_j)$ для класса "конструкции с обстоятельством места " можно получить только после прогона всех трёх моделей на реальном размеченном тестовом наборе узбекских предложений этого типа.

Ошибки необходимо разделять по причинам. Минимальная классификация включает морфологическую неоднозначность, омонимию, вариативный порядок слов, сложные конструкции, прямую и косвенную речь и случаи, которые не покрываются формальной грамматикой.

Особенно важны примеры, в которых нейронная модель выбирает неверную структуру, а формальная проверка отклоняет её. Такой пример демонстрирует реальный вклад символического компонента. Для каждого исправленного случая следует сохранять правило, которое не позволило принять ошибочный кандидат.

Разработана схема гибридного синтаксического анализа официально-деловых предложений узбекского языка, в которой нейронная и формальная подсистемы выполняют разные функции. Нейронный компонент отвечает за генерацию вероятностных кандидатов и использование контекста. Формальная грамматика ограничивает пространство допустимых структур. После проверки выполняется ранжирование, а результат сопровождается формальным объяснением.

Главное преимущество предложенной архитектуры состоит не в механическом объединении двух моделей. Сильная сторона подхода заключается в разделении ответственности: нейронная модель ищет варианты, формальная система проверяет их допустимость. Это особенно важно для официально-делового текста, где ошибка синтаксического анализа может повлиять на дальнейшую автоматическую обработку документа.Для окончательной экспериментальной версии необходимо провести обучение нейронной модели на размеченном корпусе, зафиксировать точное множество конструкций. Для математического эксперимента эти данные следует трактовать как формальное ядро, а не как готовые результаты нейронного эксперимента.

Литература

1. Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.
3. Dozat, T., & Manning, C. D. (2017). Deep Biaffine Attention for Neural Dependency Parsing. ICLR.
4. Garcez, A. d'A., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd Wave. Artificial Intelligence Review, 56, 12387–12406.

5. Jurayeva, N. V. (2020). The formal model of the grammar of the Uzbek language. European Journal of Research and Reflection in Educational Sciences, 8(4), Part II, 129–133.
6. Jurayeva, N. V., & Jurayeva, U. V. (2020). Формальные модели членов предложения грамматики узбекского языка в рамках официально-делового стиля. Universum: Психология и образование, 12(78).
7. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing. 3rd ed., draft.