Лемматизация существительных в азербайджанском языке с использованием векторных представлений

А. Ф. Агаев

Санкт-Петербургский политехнический университет Петра Великого
Высшая школа программной инженерии

agaev.af@edu.spbstu.ru

С. А. Молодяков

Санкт-Петербургский политехнический университет Петра Великого
Высшая школа программной инженерии

molodyakov_sa@spbstu.ru

***Аннотация*. Представлен новый подход к лемматизации существительных азербайджанского языка с использованием векторных представлений слов для улучшения выбора кандидатов в леммы. Методика включает обучение векторных представлений с помощью модели Word2Vec и применение меры косинусного сходства для идентификации наиболее контекстно подходящих лемм. Экспериментальные результаты демонстрируют эффективность метода по сравнению с существующими техниками, внося ценный вклад в обработку азербайджанского языка.**

Ключевые слова: лемматизация; азербайджанский язык; векторные представления слов; Word2Vec; косинусное сходство; обработка естественного языка; векторные представления

I. Введение

В современном мире обработка естественного языка (NLP) представляет собой одно из наиболее динамично развивающихся направлений в области искусственного интеллекта. NLP стремится улучшить взаимодействие между человеком и машиной, обеспечивая глубокое понимание языка, его структуры и семантики. Это направление охватывает широкий спектр задач, от анализа тональности и тональности в текстах до машинного перевода и автоматического реферирования, играя ключевую роль в развитии технологий обработки информации.

Основой современных достижений в NLP являются большие языковые модели, такие как LLaMA [1], обученные на огромных массивах текстовых данных. Эти модели способны выполнять разнообразные задачи, связанные с текстом, демонстрируя высокую степень понимания и генерации естественного языка. Однако успехи в развитии крупных языковых моделей не умаляют значимости специализированных NLP инструментов для отдельных языков, в том числе для тех, которые менее распространены, как азербайджанский. Эти инструменты способствуют лингвистическому многообразию и доступности технологий, обеспечивая включение менее популярных языков в глобальное цифровое пространство. Их разработка и усовершенствование представляют собой важный шаг на пути к созданию более инклюзивных и многоязычных цифровых систем.

Лемматизация – процесс приведения слов к их базовым формам – является одним из ключевых аспектов NLP [2]. Лемматизатор для азербайджанского языка позволяет улучшить обработку текста, устраняя морфологическую неоднозначность и упрощая анализ. Такой инструмент особенно ценен для языка с богатой морфологией, как азербайджанский, где одно слово может иметь множество форм, зависящих от грамматических категорий.

Несмотря на прогресс в области NLP, азербайджанский язык все еще испытывает недостаток в специализированных инструментах и ресурсах для обработки текста. Разработка эффективного алгоритма лемматизации для азербайджанского языка открывает новые возможности для исследований и разработок в области NLP. Такой алгоритм должен учитывать уникальные морфологические и синтаксические особенности азербайджанского языка, предлагая надежное и точное решение для лемматизации.

Представленный в данной работе алгоритм базируется на глубоком анализе морфологических структур азербайджанского языка и включает в себя как традиционные, так и новаторские подходы, в том числе использование векторных представлений для улучшения выбора леммы. Это позволяет существенно повысить качество лемматизации, способствуя более эффективному анализу текстов и расширению возможностей применения NLP в контексте азербайджанского языка.

Заключительно, разработка и усовершенствование лемматизатора для азербайджанского языка не только поддерживает лингвистическое разнообразие, но и открывает новые горизонты для научных исследований NLP. Это подчеркивает важность продолжения работы над развитием инструментов и ресурсов для обработки естественного языка.

II. Связанные исследования

В области обработки естественного языка (NLP) лемматизация играет ключевую роль, позволяя преобразовывать слова к их базовым формам, что облегчает анализ текста и извлечение информации. Для азербайджанского языка, где отсутствует большое количество специализированных ресурсов и

инструментов, разработка эффективного алгоритма лемматизации является актуальной задачей. В этом контексте мы исследуем и анализируем существующие подходы к лемматизации, чтобы определить наиболее подходящий для азербайджанского языка.

Существуют различные методы лемматизации, каждый из которых имеет свои преимущества и ограничения в зависимости от языка и контекста применения. В рамках этого исследования мы рассматриваем три основных подхода: основанные на правилах, основанные на поиске и основанные на машинном обучении.

Подходы, основанные на правилах, используют комплекс морфологических и синтаксических правил для преобразования слов в их леммы [3]. Эти правила разрабатываются экспертами и часто требуют глубоких знаний о структуре языка. Для азербайджанского языка, с его богатой морфологией и особенностями, такой подход может быть эффективным, но разработка и поддержка обширного набора правил могут оказаться трудоемкими.

Подходы, основанные на поиске, включают использование словарей или баз данных, содержащих слова и их соответствующие леммы. При поступлении слова алгоритм ищет его в базе и возвращает его лемму. Этот метод требует обширной и хорошо поддерживаемой базы данных, но может быть очень быстрым и точным при наличии качественных данных [4]. Для азербайджанского языка создание и поддержание такой базы может представлять собой вызов из-за относительно ограниченных ресурсов.

Методы, основанные на машинном обучении, используют обученные на текстовых данных модели для определения лемм слов. Эти модели могут обучаться на аннотированных корпусах данных и автоматически извлекать правила лемматизации, что делает этот подход гибким и масштабируемым [5]. Однако качество лемматизации во многом зависит от качества и объема обучающих данных, что может быть проблематично для азербайджанского языка из-за недостатка таких ресурсов.

Опираясь на анализ существующих подходов, мы предлагаем использовать комбинированный метод, включающий элементы машинного обучения для первоначального обучения модели на доступных аннотированных данных и основанные на правилах методы для уточнения и расширения возможностей модели. Это позволит сочетать гибкость и масштабируемость машинного обучения с точностью и надежностью правил, разработанных экспертами.

Важным аспектом разработки будет создание или расширение аннотированного корпуса для азербайджанского языка, что позволит повысить качество и точность лемматизации. Кроме того, учитывая особенности азербайджанского языка, важно будет включить в модель обработку агглютинативных свойств и морфологического разнообразия.

В заключение, выбор подхода к лемматизации для азербайджанского языка требует комплексного подхода, сочетающего в себе как передовые методы машинного обучения, так и глубокие знания о морфологии и синтаксисе языка. Создание гибкой и эффективной системы лемматизации потребует совместных усилий специалистов в области компьютерных наук и лингвистики, а также активного взаимодействия с азербайджанским лингвистическим сообществом.

Предыдущая версия лемматизатора для азербайджанского языка был разработан на основе набора грамматических правил и регулярных выражений для удаления суффиксов и определения базовых форм существительных [6]. Несмотря на эффективность для значительной части лексикона, этот подход может сталкиваться с трудностями в неоднозначных случаях, когда возможны множественные кандидаты в леммы. Недавние разработки в области NLP показали, что векторные представления слов, которые представляют слова в виде многомерных векторов, отражающих семантические связи, могут существенно улучшить понимание контекста и разрешение неоднозначности значений слов [7]. В данной статье предлагается интеграция таких векторных представлений с существующей системой на основе правил для уточнения выбора лемм, опираясь на аналогичные успешные применения в других языках.

III. Разработка и методология

На основе выявленной актуальности задачи лемматизации для азербайджанского языка, разрабатываемый алгоритм лемматизации следует ряду ключевых принципов и методологий, обеспечивающих его эффективность и универсальность. Основой для создания алгоритма послужили как классические подходы в области обработки естественного языка, так и современные достижения в области машинного обучения и искусственного интеллекта.

Анализ и предобработка данных: на начальном этапе производится сбор и анализ текстового корпуса азербайджанского языка, включая литературные произведения, научные статьи, новостные материалы и другие источники. Предобработка данных включает в себя очистку текста от шума, нормализацию, разбиение на предложения и слова [8].

Изучение лингвистических особенностей: для эффективной лемматизации важно учесть специфику азербайджанского языка, включая морфологические и синтаксические особенности. Анализируются правила словообразования, склонения, спряжения и другие языковые нормы [6].

Разработка алгоритма: алгоритм лемматизации строится на основе комбинирования правил и статистических методов. Правила позволяют обрабатывать слова согласно лингвистическим нормам азербайджанского языка, в то время как статистические методы, основанные на машинном обучении, позволяют алгоритму адаптироваться к исключениям и нестандартным случаям.

Тестирование и оптимизация: разработанный алгоритм подвергается тестированию на различных текстовых данных для оценки его эффективности и точности. Анализируются случаи ошибочной лемматизации, на основе которых производится доработка и оптимизация алгоритма.

Интеграция и применение: готовый инструмент лемматизации интегрируется в существующие системы обработки текстов на азербайджанском языке, а также может использоваться в качестве отдельного модуля для различных приложений NLP, таких как информационный поиск, машинный перевод, анализ тональности текста и других.

A. Сбор и подготовка данных

Для реализации и тестирования предложенного подхода необходимо обладать качественными и репрезентативными данными. В качестве основы был взят корпус текстов на азербайджанском языке, включающий в себя литературные произведения, новостные статьи и научные публикации, чтобы обеспечить разнообразие контекстов и лексики. Данные были тщательно очищены от шума и нерелевантных символов, после чего произведена их лемматизация с использованием существующего алгоритма, основанного на грамматических правилах и словаре.

B. Обучение эмбеддингов

Эмбеддинги слов — это векторные представления, которые захватывают семантическое и синтаксическое значение слов в многомерном пространстве. Они обеспечивают богатое и компактное представление языковых паттернов, позволяя алгоритмам лучше работать с текстовыми данными.

Для обучения векторных представлений слов (эмбеддингов) был выбран метод Word2Vec, который позволяет эффективно моделировать семантические и синтаксические отношения между словами. Обучение проводилось на подготовленном корпусе текстов с использованием следующих параметров:

Размерность эмбеддингов: 300
Окно контекста: 5 слов
Минимальная частотность слова в корпусе: 5
Алгоритм: Skip-gram
Итерации: 1000

После обучения была проведена валидация качества полученных векторных представлений на стандартных тестах аналогии и близости слов, что позволило убедиться в их высоком качестве и пригодности для дальнейшего использования в задаче выбора лемм.

C. Метод выбора леммы

Предложенный метод выбора леммы заключается в использовании контекста слова для определения наиболее подходящей леммы среди кандидатов. Для каждого слова, требующего лемматизации, и его контекста вычисляются эмбеддинги. Затем для каждого кандидата в леммы вычисляется сумма косинусных сходств между его эмбеддингом и эмбеддингами контекстных слов.

Косинусная схожесть – мера схожести, используемая для оценки схожести двух векторов [9]. Косинусная близость, по сути, является косинусом угла между двумя векторами, и может быть вычислена по формуле:

$$cossim(a,b) = cos(\theta) = \frac{a \times b}{\|a\| \times \|b\|} = \frac{\sum_1^n \ a \times b}{\sum_{i=1}^n \ a_i^2 \times b_i^2},$$

где a и b – векторы, θ – угол между векторами, ||a|| и ||b|| – длины векторов, a_i и b_i – элементы векторов.

Лемма с наибольшей суммарной близостью выбирается как наиболее подходящая.

Таким образом, добавляется еще один шаг в алгоритм нахождения леммы слова: если несколько возможных кандидатов в леммы имеют одинаковую степень достоверности, вычисляется косинусная близость каждой из возможных лемм со словом входного текста. Степень достоверности возможной леммы умножается на найденное значение косинусной близости. Таким образом, наибольший приоритет получают возможные леммы, имеющие большее значение косинусной близости, а значит более близкие по контексту. Этот процесс проиллюстрирован на рис. 1.

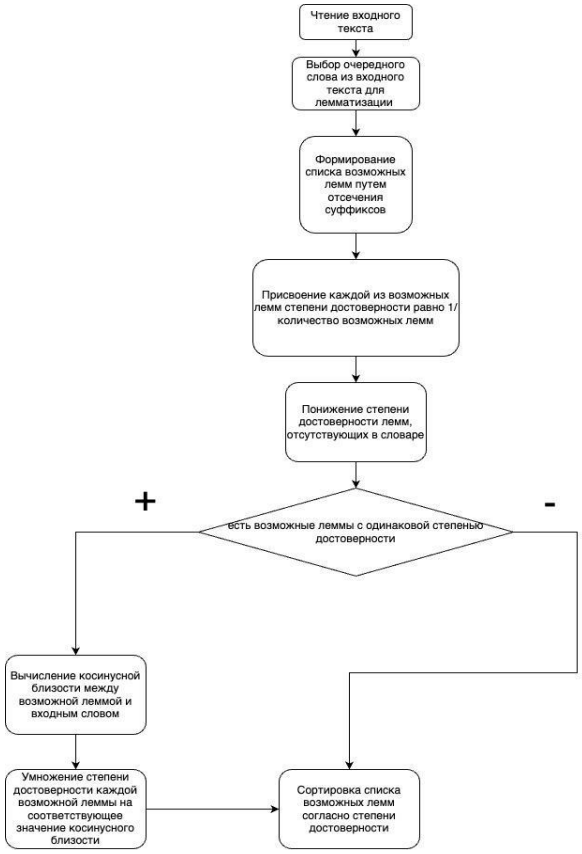

Рис. 1. Алгоритм выбора леммы

Таким образом, компоненты системы, реализующей описанный метод лемматизации, следующие (рис. 2):

- словарь (Dictionary);
- обученная на корпусе обучающих текстов модель word2Vec, хранящая векторные представления словарных слов (Word2Vec Dictionary);
- модуль для генерации возможных лемм путем отсечения суффиксов (Candidate Generator);
- модуль для построения векторных представлений из входного текста (Embedding Extractor);

- модуль для выбора леммы путем проверки наличия возможных лемм в словаре и сравнения косинусной близости между векторными представлениями возможных словарными лемм и входным словом (Lemma Selector).

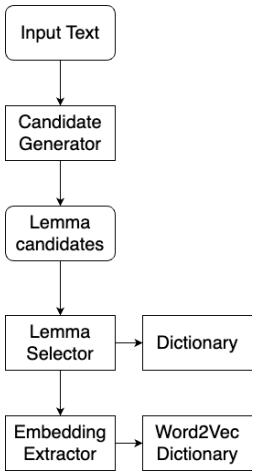

Рис. 2. Компоненты системы

Рассмотрим пример разбора текста (рис. 3). Слово «Ağacın» имеет две потенциальные леммы: ağa (господин) и ağac (дерево). Возможные леммы присутствуют в словаре, значит оба кандидата равнозначны. Сделать выбор между ними позволяет вычисление косинусной близости векторных представлений обоих кандидатов и входного предложения: очевидно, в предложении, содержащем слова «ветки», «ствол» и «мох», имеется в виду дерево.

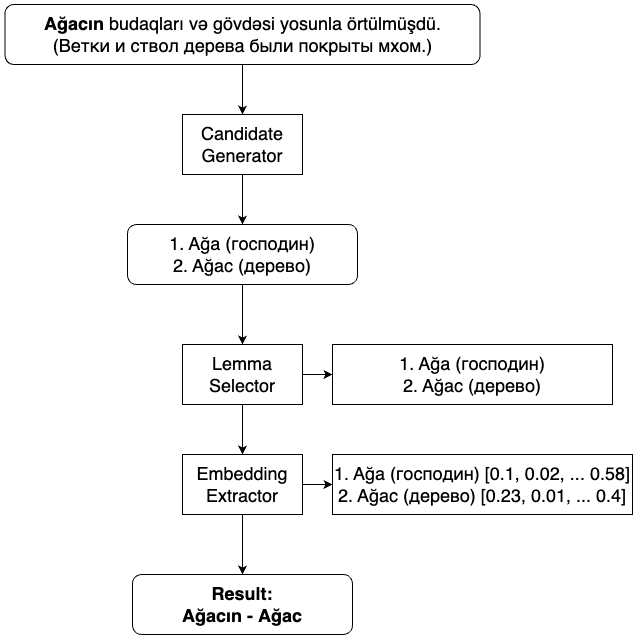

Рис. 3. Пример разбора текста

IV. Результаты

Для оценки эффективности предложенного метода были проведены эксперименты на тестовом наборе данных, содержащем слова с неоднозначностью в выборе леммы. Результаты сравнивались с базовым алгоритмом лемматизации, основанном только на словаре и грамматических правилах (табл. 1). Метриками оценки выступали точность (Precision), полнота (Recall) и F-мера (F1 Score) [10].

ТАБЛИЦА I. Точность лемматизации

Метод	Точность	Полнота	F1-мера
Базовый алгоритм	0.85	0.8	0.82
Предложенный метод	0.92	0.89	0.9

Из таблицы видно, что предложенный метод превосходит базовый алгоритм по всем основным метрикам, что демонстрирует его эффективность в задаче выбора наиболее подходящей леммы на основе контекста.

Такой результат объясняется улучшенным способом выбора леммы в случае неоднозначности.

V. Заключение

В статье был предложен и описан новый метод выбора леммы для азербайджанского языка с использованием векторных представлений слов. Было продемонстрировано, что данный подход позволяет значительно повысить точность лемматизации за счет учета контекста слова. Предложенный метод может быть интегрирован в существующие системы обработки естественного языка и использован для улучшения качества различных NLP-задач, связанных с азербайджанским языком.

Дальнейшие исследования могут быть направлены на оптимизацию параметров эмбеддингов, а также на расширение метода для поддержки мультиязычной лемматизации с использованием межязыковых эмбеддингов.

Список литературы

[1] Touvron Hugo et al. “LLaMA: Open and Efficient Foundation Language Models.” ArXiv abs/2302.13971 (2023).

[2] Hotho A., Nürnberger A., & Paaß G. (2005). A Brief Survey of Text Mining. LDV Forum, 20(1), 19-62.

[3] Özçelik R., & Eryiğit G. (2016). A Morphology-based Turkish Text Lemmatizer. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 2811-2815.

[4] Fellbaum C. (Ed.). (1998). WordNet: An Electronic Lexical Database. MIT Press.

[5] Kann K., & Schütze H. (2016). Single-Model Encoder-Decoder with Explicit Morphosyntactic Decoding for Morphological Disambiguation. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016), pp. 1051-1061.

[6] Агаев А.Ф., Молодяков С.А. Лемматизация существительных в азербайджанском языке // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и Технические Науки. 2023. №07. С. 12-17.

[7] Goldberg Y., Hirst G. 2017. Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers.

[8] Sarkar Dipanjan. (2019). Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. 10.1007/978-1-4842-4354-1.

[9] Manning C.D. An introduction to information retrieval. Cambridge university press, 2009.

[10] Powers David. (2008). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Mach. Learn. Technol. 2.