ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

Мультиязычное автоматическое извлечение лингвистической информации из грамматик

Multilingual Automatic Extraction of Linguistic Data from Grammars

Студентка 2 курса

группы № 204

Корнилова Анастасия Андреевна

Научный руководитель

Толдова Светлана Юрьевна

Доцент

Москва, 2022 г.

1. Введение

Данная работа посвящена рассмотрению методов автоматического извлечения информации из текста (Information Extraction, IE). IE является одной из подзадач обработки естественного языка (Natural Language Processing, NLP). Методы автоматического извлечения информации широко применяются для создания поисковых систем. Кроме веб-сервисов, предназначенных для поиска сайтов в Интернете, релевантных запросу пользователя, существует потребность в узкоспециализированных поисковых системах для научных публикаций, в том числе и лингвистических.

Одним из видов научных работ в лингвистике является грамматика, представляющая собой описание фонетических, морфологических, синтаксических, семантических и иных характеристик конкретного языка. До недавнего времени извлечение языковых характеристик из грамматик и создание баз данных на их основе выполнялось вручную. Например, база данных wals.com, изначально опубликованная в виде книги (Haspelmath et al. 2005), содержит информацию о 144 характеристиках для более 2600 языков.

Поиск информации о таком количестве признаков — долгий и трудозатратный процесс даже в том случае, если грамматика доступна не только в форме книги, но и в оцифрованном виде. Во-первых, грамматики из разных временных периодов (от миссионерских грамматик XVI в. до современных работ), составленные исследователями из разных стран мира, не имеют единой структуры. Во-вторых, простой поиск по слову в документе может вернуть десятки вхождений, и далеко не все из них будут релевантны запросу.

Целью данной работы является создание поисковой системы для грамматик, которая бы облегчила и ускорила процесс поиска информации о языковых характеристиках. Задачи работы — рассмотреть два метода извлечения информации (BM25 и reranking-модель на основе BERT[1]). Материалами для работы послужили 573 грамматики, представленные на Google Drive (приложение 1).

В разделе 2 будут проанализированы уже существующие работы, в которых поставлена задача автоматического извлечения информации из грамматик; в

[1] https://github.com/google-research/bert

разделе 3 будут рассмотрены сами грамматики и процесс их предобработки; в разделе 4 будут описаны использованные нами методы извлечения информации; в разделе 5 мы сравним результаты, полученные с помощью этих методов, и рассмотрим функционал созданного нами веб-приложения, содержащего поисковую систему.

2. Обзор существующих подходов

На настоящий момент предмет автоматического извлечения информации из грамматик относительно малоизучен; в рамках проектов шведской исследовательской группы Språkbanken было опубликовано несколько научных работ, в которых рассматривается методология извлечения информации из грамматик с помощью фреймовых[2] семантических парсеров (Virk et al. 2017; Virk et al. 2019; Virk et al. 2020).

Рассмотрим методологию Språkbanken на примере следующего гипотетического предложения из грамматики:

(1) *The adjectives follow the noun they qualify.* (Virk et al. 2017: 114)
'Прилагательные следуют за существительными, которые они модифицируют'.

Необходимо выбрать ответ на вопрос: 'Каков порядок следования прилагательных и существительных в языке?' из следующих значений: 'существительное-прилагательное' и 'прилагательное-существительное'.

Семантический парсер размечает его следующим образом:

- *follow* – предикат, описывающий ситуацию следования и имеющий два аргумента. Аргумент 1 и Аргумент 2 обозначают второй и первый объекты в последовательности соответственно;
- *adjectives* – Аргумент 1;
- *nouns* – Аргумент 2.

[2] Семантический фрейм – описание прототипической ситуации или отношения между объектами. Например, в ситуации торговли присутствуют участники с ролями 'продавец', 'покупатель', 'товар' и 'деньги', взаимодействующие между собой. Примеры предикатов, относящихся к ситуации торговли: *продавать, покупать, стоить, платить* (Fillmore 1976: 25).

На основе разметки семантического парсера в качестве ответа на вопрос выбирается и заносится в базу данных вариант ‘существительное-прилагательное’.

Семантические парсеры, созданные на основе размеченных корпусов текстов из Интернета, как правило, недостаточны для описания фреймов, встречающихся в грамматиках. В (Virk et al. 2020) описана работа узкоспециализированного семантического парсера для лингвистических публикаций, созданного на основе LingFN. LingFN – корпус, состоящий из грамматик на английском языке с размеченными семантическими фреймами, подробно описанный в (Malm et al. 2018). Для извлечения информации из грамматик на других языках, кроме английского, потребовалось бы создание узкоспециализированных парсеров для каждого из языков. Так как это задача, требующая значительных временных затрат и разметки большого размера данных, далее мы рассмотрим методы, не основанные на фреймовой семантике.

Более простой метод используется в (Hammarström et al. 2020): чтобы выяснить, присутствует то или иное явление в языке, подсчитывается частотность соответствующего термина в тексте грамматики (например, термина ‘masculine’ в грамматике на английском языке, чтобы выяснить, присутствует ли в описываемом языке категория мужского рода). Даже термины с низкой, но ненулевой частотностью, встречающиеся в грамматике, могут не иметь отношения к описываемому языку, если находятся в контексте отрицания: ‘в языке X [нет явления Y] | [отсутствует категория Y] | [не обнаружена категория Y]’. Пример:

(2) *Wayana lacks a category for plural (one versus more than one).* (Tavares 2006: 150)
‘В языке Ваяна отсутствует категория множественного числа (противопоставление одного объекта нескольким)’.

Исходя из распределения вхождений каждого термина в грамматики, вычисляется пороговое значение частотности. Только термины с частотностью выше пороговой являются названиями присутствующих в языке категорий.

Такой метод не требует значительных временных затрат на разметку корпусов и универсален для грамматик на любом языке, что значительно

облегчает автоматическое создание баз данных языковых характеристик.

Однако методы, описанные в (Virk et al. 2017; Virk et al. 2019; Virk et al. 2020) и (Hammarström et al. 2020), эффективны для построения баз данных о языках именно в виде таблиц, где на пересечении строки с названием языка и столбца с названием категории находится вариант ответа на вопрос (например ‘существительное-прилагательное’) или истинностное значение, указывающее на присутствие/отсутствие той или иной категории в языке.

Формат таблиц не полностью соответствует целям нашей работы, так как для полноценной поисковой системы недостаточно извлекать одно истинностное значение; важно извлечь абзац из текста грамматики, где описаны специфические особенности искомой языковой характеристики с глоссами и примерами. Поэтому в рамках данной работы было принято решение использовать методы, ранжирующие документы (абзацы) по их релевантности поисковому запросу, введенному пользователем, чтобы в ответ на запрос вернуть оригинальный абзац из грамматики.

3. Данные

Грамматики, из которых приложение извлекает информацию, представлены на Google Drive в папке Grammars (Приложение 1). Данные разделены на 13 подпапок, каждая из них отведена под отдельную группу языков.

Подразделы, представленные в базе данных:

- уральская языковая семья;
- тюркская языковая семья;
- тунгусская языковая семья;
- языки Юго-Восточной Азии (австроазиатская и тай-кадайская языковые семьи);
- тунгусо-маньчжурская языковая семья;
- языковые семьи Северной Америки;
- языковые семьи Южной Америки;
- сино-тибетская языковая семья;
- папуасская языковая семья;
- палеоазиатские языки;

- монгольская языковая семья;
- японо-рюкюская языковая семья + корейский язык;
- язык джарава, исчезнувшие языки (мероитский, этрусский), псевдо-изолированный язык (баскский), изоляты (айнский, бангери, бурушаски).

Каждая грамматика представлена в файле формата .pdf. Для каждой грамматики в таблице grammars (Приложение 2) указана метаинформация: полный путь к файлу; есть ли у файла распознанный текстовый слой (Searchable) или же он отсутствует (Not searchable); какой язык описан в грамматике; на каком языке написана грамматика.

Изначально у большинства файлов не было текстового слоя. Такие файлы были обработаны с помощью библиотеки ocrmypdf, применяющей метод оптического распознавания текста (OCR, optical character recognition) к изображениям. К названиям грамматик, для которых текстовый слой был создан с помощью библиотеки ocrmypdf, было добавлено нижнее подчеркивание и пометка “ocr”, например: “Dagur Mongolian Grammar, Texts, and Lexicon (Martin) _ocr.pdf”.

Для последующего поиска информации по грамматикам содержимое каждого файла было предобработано. Текст грамматики считывался с помощью библиотеки pdftotext и разделялся на абзацы. В качестве разделителя было взято сочетание двух пробелов. После разделения из начала и конца каждого абзаца удалялись лишние пробелы. Так как частотны случаи “разрыва” абзаца между двумя страницами, после разделения каждая пара стоящих рядом абзацев проверялась: если второй из них не начинается с заглавной буквы и/или первый не заканчивается на точку, многоточие, вопросительный или восклицательный знак, то они снова соединялись в единый абзац.

Далее каждый абзац был поделен на токены с помощью библиотеки spaCy. Была выбрана именно эта библиотека, потому что на данный момент в ней реализованы методы предобработки текста для 22 языков. В результате токенизации из текстов были удалены знаки препинания, цифры и стоп-слова. Были использованы списки стоп-слов, реализованные в SpaCy. Так для каждой грамматики был создан словарь (dictionary). В этом словаре ключами стали сами абзацы, каждый абзац представлял собой список токенов. Значениями в словаре

стали списки номеров страниц, на которых располагается соответствующий абзац. Чтобы избежать повторной обработки текста при запросе пользователя, для каждой грамматики соответствующий ей словарь с абзацами и номерами страниц сохранялся в файл формата .json в папке Grammars_Page_Numbers, чтобы поисковой алгоритм работал уже с предобработанными данными, а не с изначальным файлом формата .pdf.

После токенизации текста была проведена лемматизация: каждый токен был заменен на его лемму с помощью библиотеки spaCy. Списки абзацев с лемматизированными токенами также сохранялись в файлах формата .json в папке Grammars_Lemmas.

Пример предложения и списка токенов, получившегося из него после приведения к нижнему регистру, токенизации, удаления стоп-слов и лемматизации:

(3) *This suffix occurs with only one adverb in the data (225).* (Tavares 2006: 407)
["suffix", "occur", "adverb", "datum"]

4. Методы

4.1. Методы ранжирования абзацев по релевантности запросу

После того, как данные были поделены на абзацы и предобработаны, необходимо реализовать поисковую систему. Система должна принять запрос от пользователя, решить, какие из абзацев релевантны его запросу, и вернуть их.

Допустим, пользователь хотел бы узнать об особенностях категории множественного числа в языке аймара. Поисковая система должна вернуть ему информацию из каждой грамматики, описывающей этот язык. Чтобы алгоритм при каждом запросе пользователя не искал в таблице grammars.xslx (Приложение 2), какие файлы являются грамматиками языка аймара, был создан словарь, где в качестве ключей хранятся названия языков, а в качестве значений – списки с именами файлов с соответствующими грамматиками и названиями языков, на которых эти грамматики написаны. Словарь сохранен в файле language_files.json.

Пример ключа и значения (для чукотского языка есть всего три грамматики, две на английском и одна на французском языке):

"Chukchi": [

["Grammars/Paleosiberian/Chukutko-Kamchatkan/Chukchi, A Grammar of (Dunn).pdf", "en"],
["Grammars/Paleosiberian/Chukutko-Kamchatkan/Chukchi; Chukchee (Bogoras) (1921).pdf", "en"],
["Grammars/Paleosiberian/Chukutko-Kamchatkan/Chukchi; Parlons Tchouktche (Weinstein).pdf", "fr"]]

Теперь предстоит сравнить запрос пользователя с каждым абзацем из грамматики и определить, какие из абзацев наиболее релевантны запросу. Для подсчета релевантности в данной работе использовались алгоритм BM25 и комбинация BM25 с эмбеддингами BERT.

4.1.1. BM25

BM25 – это семейство функций, присваивающих каждому документу (в нашем случае абзацу) коэффициент релевантности искомому запросу. Мы использовали функцию, описанную в (Trotman et al. 2012) и реализованную в классе BM25Okapi библиотеки rank-bm25[3]:

$$BM25(Q,d) = \sum_{t \in Q}^{n} IDF(t) \frac{(k_i + 1) \cdot tf_{td}}{tf_{td} + k_1 \cdot (1 - b + b \cdot (\frac{L_d}{L_{avg}}))}$$

$$IDF(t,d) = log \frac{N - df_t + 0.5}{df_t + 0.5}$$

Q – запрос, введенный пользователем и состоящий из токенов t;

d – абзац, для которого определяется релевантность;

tf_{td} – количество вхождений токена t в абзац d;

df_t – количество абзацев в грамматике, в которых встречается токен t;

N – общее количество абзацев в грамматике;

L_d – количество токенов в абзаце;

L_d – среднее арифметическое количества токенов по всем абзацам;

$b = 0.75$; $k_1 = 1.5$

4.1.2. BERT

BERT[4] (Bidirectional Encoder Representations from Transformers) – модель, созданная Google в 2018 году. Помимо многих других задач автоматической

[3] https://github.com/dorianbrown/rank_bm25
[4] https://github.com/google-research/bert

обработки языка, BERT может использоваться для ранжирования документов по релевантности запросу: модель присваивает запросу и каждому абзацу из документа вектор. Чем более похожи друг на друга запрос и абзац, тем больше между ними косинусная близость. Для данной работы использовалась модель bert-base-multilingual-cased[5], поддерживающая 104 языка.

Так как создание эмбеддингов предложений с помощью BERT и подсчет косинусной близости для каждого абзаца имеет большую алгоритмическую сложность, чем BM25, для оптимизации времени работы алгоритма возможно использование комбинированного метода BM25 + реранкинг с помощью BERT (Nogueira, Cho 2019).

4.1.3. BM25 + BERT Reranking

Комбинированный метод устроен следующим образом: с помощью более простого метода ранжирования (в нашем случае BM25) из всего документа выбираются n абзацев, релевантных запросу, а затем из них с помощью более алгоритмически сложного метода (в нашем случае эмбеддер BERT) выбираются k абзацев ($k < n$).

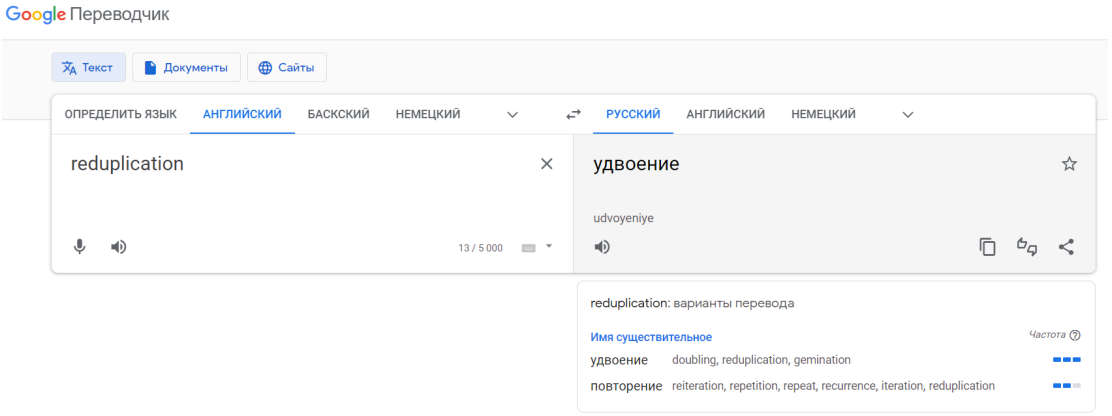
При разработке поисковой системы для грамматик было принято решение использовать только BM25 и комбинированный метод, исключив использование BERT без BM25, так как поисковая система, в отличие от алгоритмов составления баз данных, работает в реальном времени, и значительные временные задержки после ввода запроса пользователем недопустимы.

4.2. Метод решения проблемы мультиязычности

Ввиду того, что в базе данных есть грамматики не только на английском языке, для применения BM25 и реранкинга необходимо реализовать алгоритм автоматического перевода запроса пользователя с английского на другие языки. (Google Translate) и библиотеки на его основе не подходят для этой задачи: приоритет отдается наиболее частотным словам в языке, а не специфическим лингвистическим терминам. Например, лингвистическому термину *reduplication* ‘редупликация’ поставлены в соответствие два перевода: *удвоение* и *повторение*.

[5] https://huggingface.co/bert-base-multilingual-cased

Рисунок 1. (Google Translate), запрос от 25.05.2022



Следовательно, было принято решение использовать другой метод перевода лингвистических терминов на разные языки мира: с помощью Wikipedia.

HTML-код страницы Wikipedia с названием Reduplication на английском языке[6] содержит ссылки на статьи про тот же термин на других языках, например:

Русский

Метод извлечения названий статей на желаемом языке был реализован с помощью библиотеки BeautifulSoup[7]. Кроме самих названий статей, с помощью библиотеки wikipedia[8] было реализовано извлечение первых абзацев статей, где находится определение лингвистического термина. Было принято решение вместо самого переведенного на русский запроса пользователя, например, ‘*множественное число*’, подавать в качестве запроса определение:

Мно́жественное число́ (часто используется сокращение Мн., мн., мн. ч.) — грамматическое число, используемое при обозначении нескольких предметов, объединённых по какому-либо признаку (однородных предметов).

Использование определения в качестве запроса увеличивает вероятность извлечь из грамматики релевантный абзац, так как оно может содержать слова, лингвистические термины, и сокращения, которые часто встречаются в контексте

[6] https://en.wikipedia.org/wiki/Reduplication
[7] https://github.com/wention/BeautifulSoup4
[8] https://pypi.org/project/wikipedia

искомого термина: например, для запроса ‘*множественное число*’ – ‘*мн. ч.*’, ‘*грамматическое число*’, ‘*нескольких*’.

5. Обзор результатов

Рассмотрим первые пять абзацев, полученных из грамматики баскского языка (De Rijk 2007). Алгоритм BM25 возвращает 5 наиболее релевантных абзацев из каждой грамматики; в комбинированном алгоритме BM25 выбирает 10 абзацев, а затем BERT возвращает 5 наиболее релевантных из них.

Таблица 1. Сравнение BM25 и BM25 + BERT Reranking на примере грамматики (De Rijk 2007)

	BM25	BM25 + BERT Reranking
1	The first thing to note at this point is that Basque has three number categories: indefinite, definite singular, and definite plural. Any noun phrase fitted with a case ending must be morphologically marked for one of these categories.	There is, in fact, an interesting correlation between number and animacy: except when used appositively, singular dena refers to inanimates only, whereas plural denak refers to either animates or inanimates, with a definite slant toward the former. To simplify slightly, English everything corresponds to dena; English everybody to denak, requiring, of course, a plural verb form: (15) a. Dena iritsi da orain.
2	344 Chapter 15 What these examples show is the appearance of eight di¤erent forms of the auxiliary verb *edun in accord with the singular versus plural number of each of the noun phrases in the sentence.	In spoken Souletin, moreover, contrastive accent is used in order to avoid ambiguity in dealing with number categories: proparoxytonic, in case of indefinite forms, (e.g., erdimínetan) versus paroxytonic, in case of plural forms (e.g., erdiminétan).
3	Abbreviations Used in the Glosses Preliminary Remarks Following the procedure adopted by J. I. Hualde and J. Ortiz de Urbina in A Grammar of Basque (Berlin and New York: Mouton de Gruyter, 2003, pp. 7–8), no abbreviation/	These examples along with many others will be neatly accounted for by postulating the existence of an optional rule of number agreement of the perfect participle with the direct object in any syntactically transitive clause functioning as a true perfect. The agreement is

	labeling is provided, unless directly relevant, when dealing with instances to be taken as default cases of the grammatical category/categories in question. The following instances will be considered default ones (‘‘zero morphs’’ are never shown in the labeling): absabsolutive (see remarks under note 1 concerning the glossing of auxiliaries) advadverb (see note 2) artarticle auxauxiliary datdative in auxiliary glosses (see note 1) deldeletion/ellipsis In the case of morphological inflections on postpositional phrases (5.2) In the case of the following nouns: aldi (13.2.4 and 20.3), asmo (17.7.1), era (11.1.6), kontu (17.7.2), ordu (4.2, 20.2), puntu (17.7.3), uste (18.8.2), an instrumental noun to account for the meaning ‘‘cause–reason’’ in causal clauses of the -lako(tz)-type (23.1.2) In the case of the following forms: An implicit verb of saying (18.2.2 and 23.2.2), verb of subordinate -LA clauses in sentences of the type Haren begiek tximistak ziruditen (18.5.2), direlarik in an idiomatic construction used when a concession concerns an unmodified plural noun (23.7.1) ergergative in auxiliary glosses (see remarks under note 1) idfindefinite (¼ the bare stem) with proper nouns (2.3.2) and nonfinal members of an indefinite word group; with indefinite (with the exception of	680Chapter 25 implemented by the occurrence of the definite article on the participle: singular -a, plural -ak.

	beste)/ interrogative–exclamative (2.3.2)/synthetic reciprocal pronouns (15.8.1); with indefinite quantifier expressions (including bat) (2.3.2) and with locative/allative 1020Abbreviations Used in the Glosses (16.4)/privative (27.11.4) forms of the verbal noun; also indicating the bare stem of nominals, wherever appropriate/in nonfinite instrumental clauses governed by the postposition gero/geroz/geroztik (20.2.6)/in constructions with distributive numeratives (30.3) indindicative masmasculine pefperfect participle perperiphrastic conjugation prspresent ptcparticiple (This label appears in the case of perfect participle forms that do not express any shade of perfective aspect, e.g., in the combination with ezin, for which see 24.6.5.) 33rd person sinsingular definite, except for the use with perfect participle forms -sol -solidarity (characterizing 2nd-person singular) izan intransitive auxiliary *edun transitive auxiliary Remarks 1. A glossing of the type ‘‘know./plu.syt’’ in a sentence with a synthetic transitive verb indicates that there is either a singular (or indefinite, as the case may be) direct object or no such object at all and a plural subject; a gloss of the type	

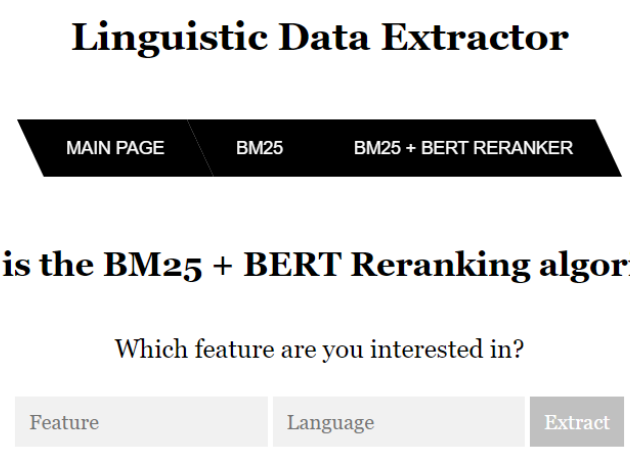
	‘‘know.plu/.syt’’ indicates a plural direct object and a third-person singular subject; in the case of ‘‘know.plu/plu.syt,’’ both direct object and subject are plural. The same method of glossing finite verb forms is used in ditransitive constructions. The relative order will always be abs–dat–erg; for example, ‘‘give.dyt’’ indicates that both the third-person subject and both objects (direct and indirect, if there is one) are singular, whereas ‘‘give.//plu.dyt’’ indicates that the subject is plural but both the objects are singular; a gloss like ‘‘give.plu/plu/.dyt,’’ however, conveys that both the objects are plural while the subject is singular.	
4	There is, in fact, an interesting correlation between number and animacy: except when used appositively, singular dena refers to inanimates only, whereas plural denak refers to either animates or inanimates, with a definite slant toward the former. To simplify slightly, English everything corresponds to dena; English everybody to denak, requiring, of course, a plural verb form: (15) a. Dena iritsi da orain.	It is true that, like the basic system, the locative system scrupulously observes the definite/indefiniteness distinction, resulting in the same three number paradigms as we had in chapter 2: definite singular, definite plural, indefinite. However, as we will soon see, it realizes this distinction in a very di¤erent fashion, marked out by a total absence of the definite article -a. As all scholars are agreed that this article developed out of a demonstrative pronoun at some relatively late period, a likely conclusion would seem to be that the locative system antedates this development and is therefore in some way more archaic than the basic system, which does involve the article -a.
5	Both terms can be used substantively to refer to persons as well as to things. In this use their semantic di¤erence is most	The first thing to note at this point is that Basque has three number categories: indefinite, definite singular, and definite plural. Any

	noticeable: batzuk: ‘some’, ‘a few’, ‘some people’, ‘a few people’; zenbait: ‘some’, ‘certain’, ‘some people’, ‘certain people’—for example, batzuen etxeak ‘a few people’s houses’; zenbaiten etxeak ‘certain people’s houses’. [N.B. Noun phrases with batzuk take plural verb forms; noun phrases with zenbait take singular or plural.] 9. In a genitive noun phrase followed by bat or batzu(k). This construction, translatable by the English expression some . . . or other, emphasizes the indefinite character of the genitivized noun phrase: katuren batsome cat or other hotsen bat some sound or other nagusiren batzuk some bosses or other alkate gazteren batzuk some young mayors or other With cardinal numbers this construction is quite common: zortziren bat sorgineight or so witches hogeiren bat urte twenty or so years More about the Basque Noun Phrase43 The reader may recall here a shorter alternative involving nominalized numerals: zortzi bat sorgin, hogei bat urte (see observation 13 of section 2.1.1). This alternative is not available for bat—a pronoun more than a numeral—in the current expression baten bat ‘someone or other’, ‘something or other’. There is also a plural form batzuren batzu(k) ‘some persons or other’, ‘some specimens or other’, although the spoken language prefers baten batzuk. 10. With certain bare nouns	noun phrase fitted with a case ending must be morphologically marked for one of these categories.

	accompanying the verbs izan ‘to be’ (see section 7.5), *edun ‘to have’ (see section 13.4.2), egin ‘to do’, ‘to make’ (see section 13.3), eman ‘to give’ (see section 15.6.6), and hartu ‘to take’, ‘to receive’ (see section 13.4.1).	

В этой грамматике BM25 в качестве самого первого абзаца выделяет стандартное описание категории множественного числа в баскском языке, чего и следовало ожидать от поисковой системы. Метод BM25 с реранкингом в качестве первого результата выделяет сведения об интересной корреляции множественного числа и одушевленности, на второе место ставит абзац о множественном числе в сулетинском диалекте баскского, а “стандартный” абзац ранжирует последним. Результаты работы двух методов варьируются на разных грамматиках. Единственная проблема заключается лишь в разделении текста на абзацы, ввиду того, что у грамматик нет единой структуры: среди результатов могут встречаться слишком длинные абзацы.

Рисунок 2. Веб-интерфейс поискового алгоритма



Пользователю предлагается выбрать алгоритм в верхнем меню, а затем ввести название языка и характеристики. Поисковая система возвращает по 5 наиболее релевантных абзацев из каждой грамматики, описывающей язык, и

после каждого абзаца выводит изображения всех страниц из pdf-файла с грамматикой, на которых этот абзац расположен. В репозитории хранится только часть грамматик; остальные грамматики, отсутствующие в репозитории, копируются с Google Drive с помощью программы rclone[9], а после выведения ответа на запрос пользователя удаляются из репозитория.

6. Заключение

В данной работе представлена система, позволяющая производить автоматический поиск релевантной лингвистической информации с автоматическим её извлечением из грамматических описаний языков, самих по себе оформленных на различных языках. Реализован веб-интерфейс для использования системы как на материалах, рассмотренных в тексте работы, так и на новых грамматиках. Проведено сравнение методов извлечения информации (классического BM25 и основанного на глубоких нейронных сетях, а также их синтеза) в задаче выделения релевантной запросу лингвистической информации на материале грамматических описаний языков. Произведена интеграция поискового алгоритма с энциклопедией Wikipedia.

Реализованная система позволяет составить впечатление об общей сложности задачи автоматизированного извлечения информации из научных текстов. Особенного внимания требует нормализация терминологии, которая варьируется в грамматических описаниях, составленных в рамках различных научных школ и эпох. Дальнейшего развития требует мультиязычность системы — тестирование произведено для грамматик, написанных лишь на крупных европейских языках, тогда как значительное количество грамматик, например, известная грамматика Пор-Рояля, составлены на иных языках (например, латыни). Быстродействие системы может быть улучшено за счет использования более быстродействующей системы, чем облако Google Drive, доступное через скрипты rclone, — например, хранилища данных в протоколе S3. Графический интерфейс может быть дополнен инструментами сбора и анализа пользовательской обратной связи.

[9] https://rclone.org

Представленная система открывает возможность для автоматических массивных типологических исследований, снимая рутинные задачи по выделению однообразной информации с исследователей-экспертов и позволяя им направить высвободившееся время на решение задач, требующих реальной экспертизы. Для дальнейшего улучшения пользовательского опыта планируется проведение тестирования системы на специалистах-лингвистах.

Литература

De Rijk 2007 — R. P. G. De Rijk. *Standard Basque: A progressive grammar.* Vol. 44. MIT Press, 2007.

Fillmore 1976 — C. J. Fillmore. Frame semantics and the nature of language. // *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, 1976. Vol. 280, No. 1, P. 20-32.

Google Translate — Google Переводчик (электронный ресурс). https://translate.google.com (Дата обращения: 25.05.2022).

Hammarström et al. 2020 — H. Hammarström, O. S. Her, M. Allassonnière-Tang. Term spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. // *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020)*, 2020. P. 27-34.

Haspelmath et al. 2005 — M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie. *The world atlas of language structures.* Oxford: Oxford University Press, 2005.

Malm et al. 2018 — P. Malm, S. M. Virk, L. Borin, A. Saxena. LingFN: Towards a framenet for the linguistics domain. //*11th edition of the Language Resources and Evaluation Conference*, *7-12 May 2018, Miyazaki (Japan)*, 2018. P. 37-43.

Nogueira, Cho 2019 — R. Nogueira, K. Cho. Passage Re-ranking with BERT. // *arXiv preprint arXiv:1901.04085.* 2019.

Tavares 2006 — P. S. Tavares. *A grammar of Wayana*. Diss. Rice University, 2006.

Trotman et al. 2012 — A. Trotman, X. Jia, M. Crane. Towards an Efficient and Effective Search Engine. // *OSIR@ SIGIR*. 2012. P. 40-47.

Virk et al. 2017 — S. M. Virk, L. Borin, A. Saxena, H. Hammarström. Automatic extraction of typological linguistic features from descriptive grammars. // *International Conference on Text, Speech, and Dialogue.* Cham: Springer, 2017. P.

111–119.

Virk et al. 2019 — S. M. Virk, A. S. Muhammad, L. Borin, M. I. Aslam, S. Iqbal, N. Khurram. Exploiting frame-semantics and frame-semantic parsing for automatic extraction of typological information from descriptive grammars of natural languages. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 2019. P. 1247–1256.

Virk et al. 2020 — S. M. Virk, H. Hammarström, L. Borin, M. Forsberg, S. K. Wichmann, M. Ionov, J. P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil, J. Gracia. From linguistic descriptions to language profiles. // *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, 2020. P. 23–27.

Приложение

Приложение 1:
https://drive.google.com/drive/folders/1-GRhRW7KyluVOwTmoln9C_5bIyzgSLeN?usp=sharing

Приложение 2:
https://docs.google.com/spreadsheets/d/1G_wbpzBb-4NUzhOQ-eYwvXTJKRbGDavClNKsSVxDeZ0/edit?usp=sharing

Исходный код поисковой системы:
https://github.com/entrapolarity/linguistic-data-extraction