OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ВЕБ-ИНТЕРФЕЙС ДЛЯ СНЯТИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ В КОРПУСЕ ТАТАРСКОГО ЯЗЫКА

Гильмуллин Р.А. Гатауллин Р.Р.,

Казанский Федеральный университет,
Институт «Прикладной семиотики» Академии Наук Республики Татарстан,
г. Казань, Российская Федерация

rinatgilmullin@gmail.com

ramil.gata@gmail.com

В работе описывается веб-инструментарий, который предназначен для ручного снятия морфологической многозначности в корпусе татарского языка. Отмечены цели и задачи проекта. Описан основной функционал инструментария.
Ключевые слова: разрешение морфологической многозначности, электронный корпус языка, татарский язык.

Введение

Для решения многих лингвистических задач используются электронные коллекции текстовых документов, так называемые электронные корпуса. При этом наиболее информативными и полезными являются размеченные корпуса, в которых текстовым единицам приписана лингвистическая информация. Обычно под такой информацией подразумеваются морфологические, синтаксические и семантические характеристики языка. Корпус может быть использован как в решении исследовательских задач, так и в разработке прикладных приложений, использующие лингвистические модели языка.

В настоящее время специалистами научного-исследовательского института «Прикладная семиотика» Академии наук Республики Татарстан активно ведутся работы по созданию электронного корпуса татарского языка «Туган тел» [Сулейманов и др., 2011]. В данный момент корпус содержит порядка 40 млн. словоформ. Осуществлена полная автоматическая морфологическая разметка корпуса [Сулейманов и др., 1997].

Недостатком данной автоматической разметки является анализ слова в отрыве от контекста, порождая в виде результата всевозможные варианты разбора слова, только один из которых актуален в конкретном контексте. Таким образом, появляется морфологическая многозначность, разрешение которой является следующим этапом в разметке корпуса.

Согласно анализу данных морфологической разметки, доля слов с морфологической многозначностью в корпусе текстов составляет порядка 31,33% [Хакимов, 2014]. В основном к проблеме разрешения морфологической многозначности в татарском языке относится проблема разрешения функциональной омонимии. Функциональная омонимия – омонимия, когда слова совпадают в написании лишь в определенных формах, являясь при этом разными частями речи.

Как показывает анализ существующих методов, проблема снятия морфологической многозначности решалась исследователями разными способами. Первые алгоритмы были основаны на правилах. Позже для решения задачи элиминации многозначности начали применяться статистические алгоритмы. У каждого подхода есть как преимущества, так и недостатки. В условиях татарского языка применение лишь одного подхода не представляется до конца возможным, вследствие чего предлагается применять некий сплав методов, перекрывая недостатки одного метода преимуществами другого [Гатауллин и др., 2014] [Гатауллин, 2014][Зинькина и др., 2005].

Как для статистического подхода нужна обучающая выборка, так и для разработки контекстных правил необходима вручную размеченная часть корпуса со снятой многозначностью для последующего анализа. Вследствие чего было принято решение о разработке инструментария для удобной разметки корпуса, которая в первую очередь будет использоваться как обучающая выборка для

машинного обучения методов разрешения многозначности.

1. Требования к системе

Для правильной работы методов, основанных на статистическом подходе, необходимо иметь достаточно большой объем обучающей выборки, охватывающий всю предметную область. В случае татарского языка, в котором количество условных типов многозначности превышает 10.000 для корпусной выборки в 21 млн. словоупотреблений, появляется необходимость в достаточном количестве разрешенных контекстных ограничений для каждого конкретного типа. Этим обусловливается необходимость охвата большой аудитории, привлечение как можно больше участников.

Однако при увеличении количества участников, появляется проблем несогласованности действий, когда одно и то же действие бессмысленно выполняться несколькими пользователями. При этом теряется эффективность и тратится время.

Также для разрешения такого рода многозначности не обязательно быть профессиональным лингвистом, достаточно знать язык на хорошем уровне. Но это не освобождает от вопроса проверки компетентности участников и верификации результатов.

Данные проблемы должны были так или иначе быть решены. Кроме всего прочего инструментарий должен быть надежным, удобным и простым в использовании.

2. Функциональность

В итоге, анализируя опыт исследовательских коллективов, занятых в этой области [Бочаров и др., 2011], было принято решение разработать веб-интерфейс для ручного снятия многозначности в корпусе. Во-первых, это мотивировалось тем, что инструментами, доступными через всемирную сеть Internet, намного легче привлечь большое количество людей. Во-вторых, вся информация по результатам снятия многозначности хранится не разрозненно, а в одном месте. И при необходимости их можно сравнить (для верификации результатов). В-третьих, веб-инструментарий разработан таким образом, что можно согласовать некоторые действия пользователей, например, давать на разрешение только определенные типы омоформы, приоритет которых на данный момент выше, чем у других типов и т.д.

Как и было отмечено ранее, основной функционал инструмента заключается в ручном снятии морфологической многозначности в конкретном контексте, т.е. каждому конкретному зарегистрировавшемуся участнику предлагается конкретное предложение из корпуса с конкретным словом с несколькими возможными разборами для определения правильного в данном контексте варианта.

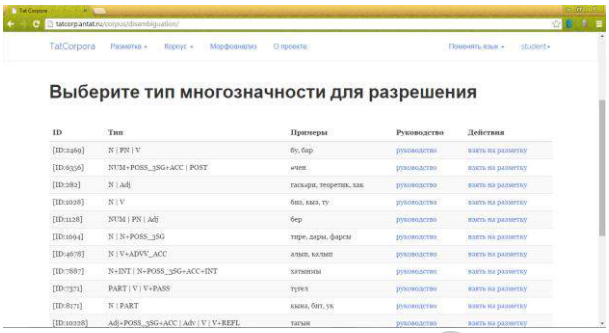

Рисунок 1 – Режим «разрешения по типам». Окно выбора типа многозначности для разбора.

Приложение поддерживает два режима функционирования: разрешение по типам и разрешение по предложения. Первый режим предполагает разрешение только определенных типов многозначности, приоритет которых на данный момент выше, чем у других типов (см. Рис.1). Например, это могут быть самые частотные типы или интересные и значимые с точки зрения лингвистики типы многозначности. Данный режим также предусматривает наличие руководства для разрешения каждого предложенного типа, что в какой-то степени должно облегчит процесс разрешения для участников.

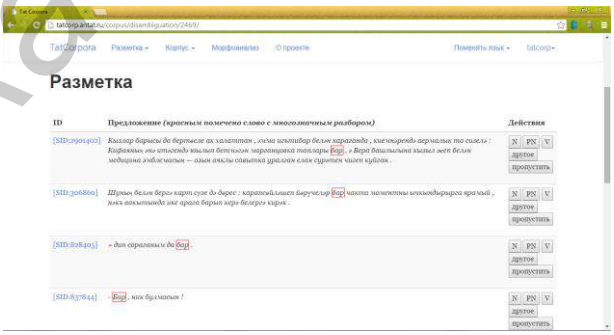

Рисунок 2 – Режим «разрешения по типам». Окно разметки. Кроме возможных типов есть варианты «пропустить» и «другое».

Предполагается, что результаты разрешения будут использоваться для обучения статистических методов разрешения, вследствие чего для максимального покрытия многозначных слов для разрешения будут выбираться самые частотные типы, а количество уникальных контекстов для каждого типа будет ограничено достаточным для приемлемого обучения методов количеством, например, 1000 примеров по конкретному типу.

Также предусмотрена верификация результатов разрешения, сравнивая результаты нескольких пользователей и выбирая только те варианты, в которых мнение большинства пользователей сошлось. Со временем, исходя из результатов разрешения, можно помечать уровень компетентности участников, тем самым повышая качество разрешения.

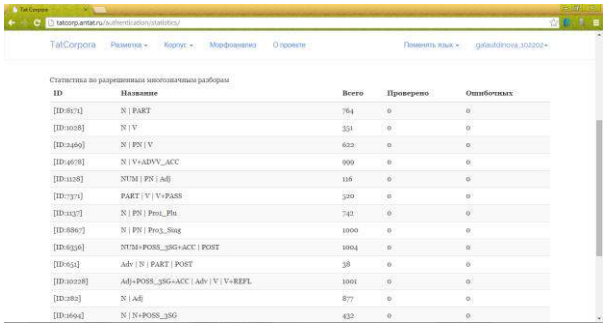

Рисунок 3 – Окно «Моя статистика» отображает полную статистику по разрешениям для данного конкретного пользователя.

Второй режим предполагает разрешение многозначностей по предложениям. Данная функция больше приурочена к подготовке полностью размеченного корпуса со снятыми многозначностями.

Также каждый зарегистрированный пользователь может следить за результатами своих разрешений во вкладке «Моя статистика» (см. Рис.3), а также сравнивать свои результаты с результатами других пользователей, тем самым включается соревновательный момент, немаловажный пункт в «краудсорсинговых» (англ. crowdsourcing, crowd — «толпа» и sourcing — «использование ресурсов») приложениях.

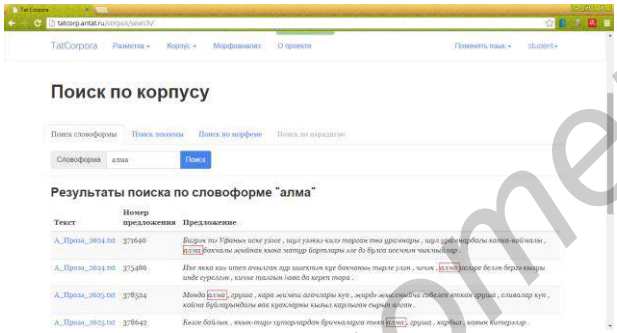

Рисунок 4 – Корпус-менеджер. Окно поиска с вкладками «Поиск по словоформе», «Поиск по лексеме», «Поиск по морфеме».

Из дополнительного функционала можно отметить функцию довольно простого корпус-менеджера (см. Рис.4). Реализованы поиск по словоформе, поиск по лексеме, поиск по морфеме и поиск по типу морфологической многозначности.

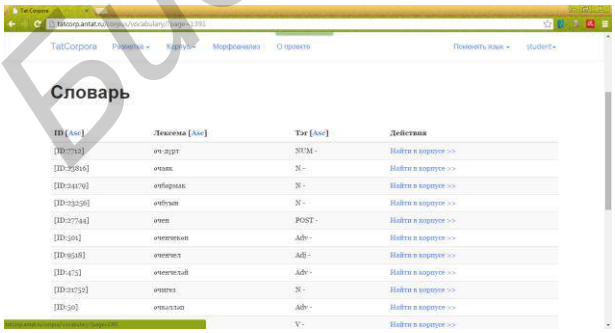

Рисунок 5 – Корпус-менеджер. Окно словаря. Реализован «поиск по лексеме».

Также на сайте присутствует информация о типах многозначностей, примеры по ним, и количество экземпляров данного типа, представленного в корпусе татарского языка (см. Рис.6). Данная информация может быть полезна исследователям, занимающимся данной проблематикой.

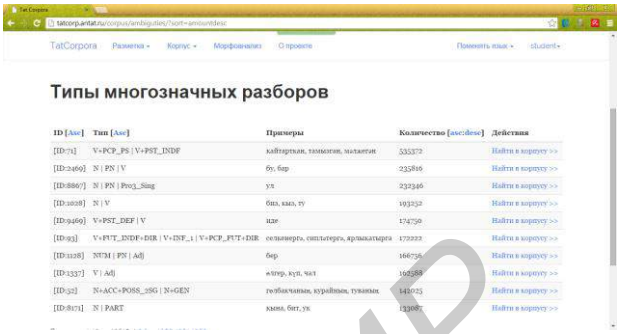

Рисунок 6 – Корпус-менеджер. Окно «Типов морфологических многозначностей». Реализован «поиск по типу многозначности».

Данный инструментарий разрабатывался, используя открытые программные обеспечения, такие как СУБД Postgresql для управления базами данных и веб-фреймворкDjango (язык программирования Python) для серверной части инструмента.

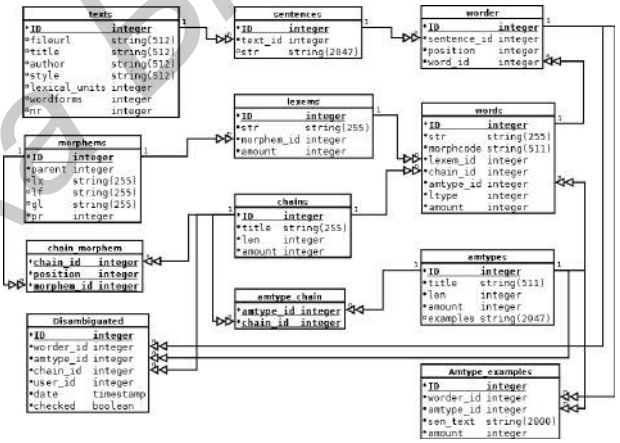

Рисунок 7 – Структура базы данных корпуса.

Заключение

В настоящее время разработан веб-интерфейс, который реализован в виде краудсорсинговой платформы с возможностью привлечения большего количества участников проекта для снятия морфологической многозначности в корпусе татарского языка. На сайте зарегистрировано 32 пользователя, которые являются студентами отделения татарской филологии и межкультурной коммуникации им. Г.Тукая Института филологии и межкультурной коммуникации Казанского Федерального университета. За пять месяцев функционирования сайта суммарное количество разрешенных контекстов достигло 97471, из них уникальными являются 29768 контекста, классифицированных по 22 типам морфологическим многозначностям.

Следующим этапом работ является развитие веб-интерфейса за счет интеграции контекстных правил для автоматического снятия морфологической

многозначности [Гатауллин и др., 2014a], с возможностью непосредственного тестирования правил на корпусных данных.

Необходимо также продвижение проекта через социальные сети для привлечения большего количества участников.

Библиографический список

[Сулейманов и др., 2011] Сулейманов, Д.Ш. Корпус татарского языка: концептуальные и лингвистические аспекты / Д.Ш. Сулейманов,Б.Э. Хакимов, Р.А. Гильмуллин // Вестн. ТГГПУ. – 2011. - № 4 (26). – С.211-216.

[Сулейманов и др., 1997] Сулейманов, Д.Ш. Двухуровневое описание морфологии татарского языка / Д.Ш. Сулейманов, Р.А. Гильмуллин // Тезисы Международной научной конференции "Языковая семантика и образ мира". Казань: Изд-во Казан. гос. ун-та., 1997. Книга 2. С. 65-67.

[Хакимов и др., 2014] Хакимов, Б.Э. Разрешение грамматической многозначности в корпусе татарского языка /Б.Э. Хакимов, Р.А. Гильмуллин, Р.Р. Гатауллин // Ученые записки Казанского университета (в печати).

[Гатауллин и др., 2014a] Гатауллин Р.Р. Программный инструментарий для разрешения морфологической многозначности в татарском языке / Р. Р. Гатауллин, Д. Ш. Сулейманов, Р. А. Гильмуллин // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2014 OpenSemanticTechnologiesforIntelligentSystems МАТЕРИАЛЫ IV МЕЖДУНАРОДНОЙ НАУЧНО-ТЕХНИЧЕСКОЙ КОНФЕРЕНЦИИ (Минск, 20-22 февраля 2014 года), - Минск. : БГУИР, 2014. - С. 503-508

[Гатауллин, 2014b] Гатауллин Р. Р. Веб-инструментарий для снятия морфологической многозначности в текстовом корпусе татарского языка / Р. Р. Гатауллин // Сохранение и развитие родных языков в условиях многонационального государства: проблемы и перспективы: материалы V Международной научно-практической конференции (Казань, 19-22 ноября 2014 г.). – Казань: Отечество, 2014. - С. 71-73

[Зинькина и др., 2005] Ю.В. Зинькина, Н.В. Пяткин, О.А. Невзорова, Разрешение функциональной омонимии в русском языке на основе контекстных правил. // Труды межд. конф. Диалог'2005.– М.: Наука, 2005. С. 198-202.

[Бочаров и др., 2011] Бочаров, В. В. Программное обеспечение для коллективной работы над морфологической разметкой корпуса / В. В. Бочаров, Д. В. Грановский // Труды международной конференции «Корпусная лингвистика – 2011». 27–29 июня 2011 г., Санкт-Петербург. — СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011.

WEB-SITE FOR HANDY MORPHOLOGICAL DISAMBIGUATION IN TATAR LANGUAGE CORPUS

Gilmullin R.R., Gataullin R.R.

Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan Federal University, Kazan, Russia

rinatgilmullin@gmail.com

ramil.gata@gmail.com

Paper presents the easy in use instrument for handy elemination of morphological ambiguities in Tatar language corpus. Talks about main tasks. Shows basic functionalities.

Introduction

Language corpora are very usefull informational structures in NLP. For successful uses it must be correctly annotated, including morphological, syntactical and semantical features.

For last years, scintists from Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences have been developing Tatar language corpus, which contains by these days more than 40 million word usages. Morphological features are automatically annotatted, but the problem of morphological ambiguity has not been solved yet. For the further researches it is necessary to get the part of corpus, annotated by hand. In the end it will be used to teach the autoamtic morphological disambiguitor based in machine learning.

Main Part

Since the corpus contains very big amount of texts, it is neccesary to involve into disambiguation process as many people as possible. Nowadays, it is much easier, if we will use Internet technologies for this task. So, the purpose was to develope the easy in use web-site, where the every could help to annotate the corpus.

As a result, the instrument functions now. Except the main functionality of handy disambiguation, simple corpus manager is also developed. And it is available by Internet at http://tatcorp.antat.ru/.

Conclusion

For 5 months of functioning, 32 users had been registrated, most of them attends Kazan Federal university, and 29768 unique context had been disambiguated.

Future plan is a promotion of the project through social social network.