OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

КОНТЕКСТНЫЕ ПРАВИЛА ДЛЯ РАЗРЕШЕНИЯ 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

Данная работа является продолжением работ по подготовке размеченного корпуса татарского языка. Ранее был представлен проект (http://tatcorp.antat.ru) по ручному снятию морфологической многозначности в корпусе татарского языка, основной целью которого является использование краудсорсингового подхода для накопления данных. Следующим шагом стала разработка инструментария для создания, тестирования и апробации контекстных правил разрешения морфологической многозначности.
Ключевые слова: корпус татарского языка, разрешение морфологической многозначности, контекстный метод.

Введение

Татарский язык относится к агглютинативному типу языков, т.е. словоформы в татарском языке образуются последовательным присоединением аффиксов к основе слова. Аффиксы имеют жесткий порядок следовании, но некоторые последовательности могут повторяться, усложняя смысл словоформы и образуя таким образом теоретически бесконечно длинные словоформы (например, «урманнардагылардагылар» - те, что (кто) у тех, кто в лесах) [Сулейманов и др., 1997]. Но на практике, по статистическим данным корпуса, длина аффиксальной цепочки в среднем не превышает 5-6 аффиксов, а максимальная длина аффиксальной цепочки равна 12. Но и такая ситуация приводит к большому разнообразию типов морфологической многозначности.

Из предварительного анализа корпуса татарского языка выявлено, что почти треть всех словоформ (~31%) корпуса в той или иной мере имеют более одного разбора [Хакимов и др., 2014]. Количество типов морфологической многозначности превышает 10000 для корпусной выборки в 21 млн. словоупотреблений [Гатауллин, 2014], что вместе с агглютинативностью татарского языка теоретически приводит к бесконечному многообразию таких типов. Практически же должна быть возможность свести их к конечным классам, для которых правила разрешения будут одними и те же. На данный момент было выявлено порядка 400 таких классов, но данная гипотеза еще требует подтверждения и дальнейших исследований.

Для автоматического разрешения морфологической многозначности предлагается применять гибридный метод, включающий метод, основанный на контекстных правилах, и статистико-вероятностный метод. Такой выбор обусловливается, тем, что, несмотря на то, что метод контекстных правил показывает достаточно хорошую точность при разрешении, сама по себе разработка таких правил достаточно трудоемкая задача, требующая тщательного лингвистического анализа контекстных ограничений для каждого типа многозначности. Статистико-вероятностные методы и методы машинного обучения хорошо выявляют и успешно используют при работе скрытые закономерности и взаимосвязи в контексте, но минусом их является плохая обучаемость на разреженных данных. Для некоторых типов, действительно, в корпусе имеется мало примеров.

Таким образом, предлагается перед использованием статистико-вероятностного метода, как первоначальный этап, применять контекстный метод. Также кроме случаев с разреженными данными, контекстный метод подходит для случаев, когда многозначность возникает вследствие избыточности словаря основ морфоанализатора, и других исключительных случаев, когда легче прописать контекстные ограничения, чем готовить обучающую выборку по данному типу или случаю [Хакимов и др., 2014].

В настоящее время подготовлено веб-приложение для разметки корпуса татарского языка. Основной упор делается на «кроудсорсинговый» аспект, т.е. использование усилий большого количества людей. С помощью данного инструмента, во-первых, будет подготовлен вручную размеченный подкорпус. Во-вторых, будет получено достаточное количество данных для обучения статистико-вероятностного метода. Следующим шагом в развитии этого проекта стал инструментарий разработки контекстных правил для разрешения, использующий ранее полученные данные для тестирования и апробации разрабатываемых правил.

1. Ключевые понятия и архитектура инструментария для разработки контекстных правил разрешения морфологической многозначности

В ранних публикациях [Гатауллин и др., 2014] описывались идея и архитектура инструментария, представлен прототип, который показал работоспособность. В текущей работе инструментарий был доработан и реализован в виде веб-приложения. В работе [Зинькина и др., 2005] подробно описаны основные достоинства и недостатки метода контекстных правил, приведены конкретные структуры обобщенных правил для разрешения функциональной омонимии некоторых типов.

Обобщенный метод контекстного разрешения функциональной омонимии для татарского языка включает несколько этапов [Гатауллин и др., 2014]:

1. построение полной классификации типов функциональных омонимов;
2. выделение минимального множества разрешающих контекстов для каждого типа. Минимальность множества означает, что для каждого типа функционального омонима следует оценить сложность распознавания каждой части речи, принадлежащей данному типу. Затем необходимо построить множество разрешающих контекстов (МРК), имеющих минимальную сложность распознавания. В алгоритмической записи данное требование выражается следующим правилом: если для функционального омонима X, имеющего тип T1 или T2, подошло правило из МРК, то тип омонима X определяется примененным правилом, иначе приписывается альтернативный тип;
3. построение управляющей структуры обобщенного правила, обеспечивающего максимальную точность распознавания.

Для решения поставленных задач разработана соответствующая архитектура программного инструментария, включающая следующие базовые объекты и понятия:

- Омоформа (или функциональный омоним) – слова, совпадающие в своем звучании лишь в отдельных формах (той же части речи или разных частей речи);
- База типов омоформ (или База контекстных правил) – иерархически упорядоченный список типов омоформ; для каждого типа определено множество разрешающих контекстов. На основе этих правил происходит разрешение многозначности для отдельно взятого типа омоформ;
- Обобщенное правило разрешения (ОПР) – правило, на основе контекстной информации определяющее актуальный вариант структуры омоформы. Для каждого типа функционального омонима следует оценить сложность распознавания каждой части речи, принадлежащей данному типу;
- Множество разрешающих контекстов (МРК) – совокупность минимальных разрешающих контекстов, достаточных для распознавания функционального омонима как определенного варианта структуры омоформы;
- Управляющая структура обобщенного правила обеспечивает и контролирует порядок применения правил;
- Минимальный разрешающий контекст – неделимое в данном контексте простое условие, имеющее минимальную сложность распознавания.

Процесс распознавания омонимии происходит следующим образом:

1. У анализируемого слова определяется тип функциональной омонимии, и в соответствии с этим типом из Базы контекстных правил находится обобщенное правило разрешения;
2. Управляющая структура задает порядок применения правил;
3. При применении каждого правила, проверяется каждый минимальный контекст разрешении этого правила;
4. Если при проверке правила получили подтверждение о его истинности, то функциональная омонимия распознается в соответствии с этим вариантом структуры омоформы;
5. Иначе, если есть другое правило, осуществляется переход к следующему правилу и выполняется то же самое;
6. И если нет другого правила, то в качестве структуры выбирается тип по умолчанию;
7. Если нет такого типа, то многозначность помечается как неразрешенная.

2. Кроудсорсинговый аспект приложения

В некоторых источниках «краудсорсинг» (англ. crowd – толпа, народ; source - ресурс) трактуется как мобилизация ресурсов людей посредством информационных технологий с целью решения задач, стоящих перед бизнесом, государством и обществом в целом [5]. Действительно, для решения некоторых задач такой подход полностью оправдывает себя. Примером служат всемирно-известный ресурс Википедия, программистский

ресурс http://stackoverflow.com/ и другие разного рода форумы, где сбором информации и наполнением сайта занимаются обычные пользователи. В сфере NLP можно отметить Открытый корпус русского языка http://opencorpora.org/, где с помощью пользователей происходит разметка корпуса [Бочаров и др., 2013]. Как уже отмечалось ранее [Гатауллин, 2014], у нас также имеется опыт в таком роде проекте: с помощью пользователей ресурса http://tatcorp.antat.ru разрешается морфологическую многозначность в корпусе татарского языка.

Основная идея подхода состоит в разбиении задачи на мелкие подзадачи, которые достаточно легко решаются и не сильно затрудняют пользователя. Другой важной частью является мотивация пользователей. Для одних это всевозможные “ачивки” (англ. achievement достижение), для других развитие opensource проектов (англ. open - открытый; source - ресурс).

Применения данного подхода для разработки контекстных правил вполне возможно, но в отличие от случая ручного снятия морфологической многозначности, где требуется простое знание языка, разработка контекстных правил требует определенных знаний в области языкознания и лингвистики, что сильно ограничивает круг возможных пользователей. Но несмотря на это, есть возможность привлечения для работы студентов-лингвистов и учителей, которые занимаются данной проблематикой.

3. Разработка контекстных правил

Для того, чтобы начать разработку контекстных правил необходимо зарегистрироваться на сайте http://tatcorp.antat.ru и перейти на вкладку http://tatcorp.antat.ru/disam/rules/, где представлен список всех омоформ, для которых уже имеются правила разрешения. Имеется возможность как улучшать уже имеющиеся правила, так и создавать новые правила для не имеющихся в списке типов омоформ (см. Рис.1).

Так как в разработке участвует не один пользователь, появляется необходимость разграничения доступа пользователей, а также необходимость своего рода “песочницы”, где происходит разработка и тестирование правил. После этого правило может быть добавлено в основную базу правил.

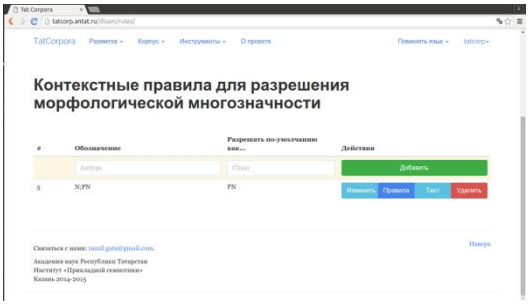

Рисунок 1 – Основная страница веб-приложения по разработке контекстных правил.

Исходя из этих соображений, был приняты простые правила:

- нельзя удалять и редактировать правила других пользователей. При желании доработать правило, нужное правило дублируется с припиской к текущему пользователю и все улучшения делаются там;
- пока правило не добавлено в основную базу правил, она находится в зоне “песочницы”, где она может редактироваться и тестироваться;
- перед добавлением (либо обновлением) в основную базу, правило тестируется на корпусных данных, и при условии успешного прохождения тестов, правило добавляется в основную базу и может быть применено в процессе разрешения многозначности.

Как уже было описано ранее, разработка правил состоит из нескольких этапов:

- сначала выбирается тип омоформ, для разрешения которых разрабатываются правила;
- Потом выбирается управляющая структура, т.е. в процессе лингвистического анализа выявляются множества минимальные разрешающие контексты и упорядочиваются по частотности так, что самые частые случаи будут рассматриваться первыми;
- для каждого множества минимальных контекстов определяется порядок минимальных контекстов (см. Рис.2).

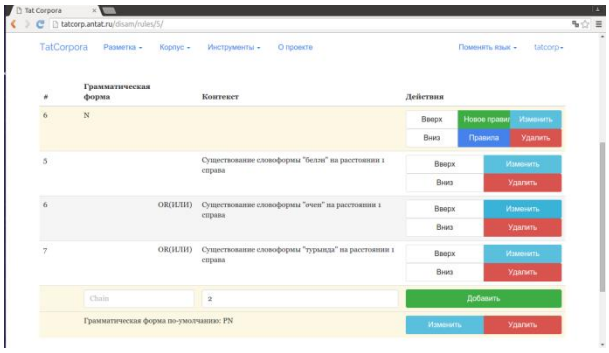

Рисунок 2 – Страница управляющей структуры правила с определенным множеством минимальных контекстов.

Кроме самой разработки имеется возможность тестирования правил на корпусных данных (см. Рис. №3). Очевидно, таким образом, легко выявляются исключительные и ошибочные случаи

распознавания, что намного облегчает процесс разработки.

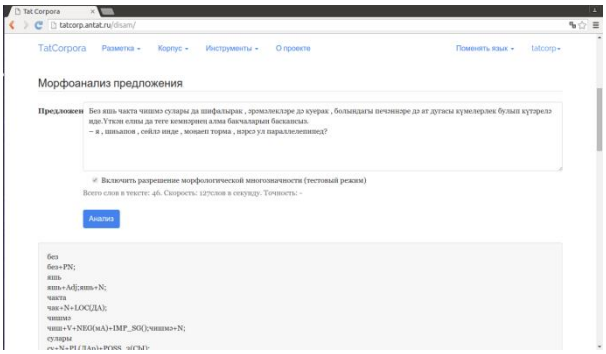

Рисунок 3 – Страница тестирования контекстных правил.

Заключение

В данной работе представлен инструментарий разработки контекстных правил для разрешения морфологической многозначности в корпусе татарского языка. При разработке особый упор делается на “краудсорсинговый” аспект приложения. На данный момент приложение на этапе тестирования. В разработке участвуют 4 человека, было разработано 9 тестовых правил, которые по большей части покрывают исключительные случаи морфологической многозначности.

Следующим шагом планируется реализация статистико-вероятностных методов для разрешения и компоновка их с методом контекстных правил.

Библиографический список

[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] Разрешение грамматической многозначности в корпусе татарского языка / Б.Э.Хакимов, Р.А.Гильмуллин, Р.Р.Гатауллин // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. - 2014. - Т. 156, кн. 5. - С. 236-244.

[Гатауллин, 2014] 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

[Гатауллин и др., 2014] 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 Open Semantic Technologies for Intelligent Systems МАТЕРИАЛЫ IV МЕЖДУНАРОДНОЙ НАУЧНО-ТЕХНИЧЕСКОЙ КОНФЕРЕНЦИИ (Минск, 20-22 февраля 2014 года), - Минск. : БГУИР, 2014. - С. 503-508.

[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., 2013] Crowdsourcing morphological annotation / Bocharov V.V., Alexeeva S.V., Granovsky D.V., Protopopova E.V., Stepanova M.E., Surikov A.V. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая–2 июня 2013 г.). Вып. 12 (19). — М.: РГГУ, 2013.

CONTEXTUAL RULE METHOD FOR MORPHOLOGICAL DISAMBIGUATION IN THE TATAR LANGUAGE

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

The article describes a software tool for creating, editing, and testing contextual rules for the automatically resolve of morphological ambiguity in the Tatar language.

Introduction

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. Since Tatar language is one of agglutinative languages, types of morphological ambiguities are theoretically infinite. It means that machine learning algorithms will not cover all cases of them, due to data sparseness. So it is necessary to combine them with rule base methods for these cases. And this work introduces such tool for Tatar language.

Main Part

Actually, the rule development tool was constructed as web service. It is available at http://tatcorp.antat.ru. To get more efficiency, “crowdsource” approach is used. It means, that rules are created with help of many users of systems. Of course, as production rules will be used only successfully tested rules.

Conclusion

For now the tool is in testing phase. By January 2016, nine testing rules were created and tested on corpus data. Next step will be the development of machine learning algorithms, which will be combined with this tool.